Evaluation automatique des contenus éducatifs en ligne basée sur l'analyse de l'apprentissage

Yosra Mourali

Thèse de doctorat

Pour obtenir le grade de Docteur de

L'UNIVERSITÉ POLYTECHNIQUE HAUTS-DE-FRANCE

et l'INSA Hauts-De-France

en Informatique et applications

et l'UNIVERSITÉ DE SFAX

en Informatique

Présentée et soutenue par Yosra MOURALI.

Le 14/12/2022, à Valenciennes

Écoles doctorales :

École Doctorale Polytechnique Hauts-de-France (ED PHF n°635)
Ecole Doctorale en Economie, Gestion et Informatique (EGI)

Unités de recherche :

Laboratoire d'Automatique, de Mécanique et d'Informatique Industrielles et Humaines (LAMIH – UMR CNRS 8201)
Laboratoire de Recherche en Technologies de l'Information et de la Communication et Génie Electrique (LaTICE - LR11ES04)

Evaluation automatique des contenus éducatifs en ligne basée sur l'analyse de l'apprentissage

JURY

Président du jury :

- Moussa Faouzi. Professeur en informatique. Université de Lorraine, France.

Rapporteurs :

- Abel Marie-Hélène. Professeur en informatique. Université de Technologie de Compiègne, France.
- Cheniti-Belcadhi Lilia. Maître de Conférences en Informatique. Université de Sousse, Tunisie.

Examinateurs :

- Ayachi Ghannouchi Sonia. Professeur en Informatique de gestion. Université de Sousse, Tunisie.

Co-directeurs de thèse :

- Jemni Mohamed. Professeur en informatique. Université de Tunis, Tunisie.
- Kolski Christophe. Professeur en informatique. Université Polytechnique Hauts-de-France, France.

Membres invités :

- Agrebi Maroi. Docteur en informatique. Société CapHornier, France.
- Farhat Ramzi. Maître-Assistant en Informatique. Université de Tunis, Tunisie.

Remerciements

Je remercie Monsieur Mohamed Jemni, Monsieur Christophe Kolski, et feu Monsieur Houcine Ezzedine, mes directeurs de thèse de m'avoir mis, dès le début, sur la bonne voie et assuré le suivi de mon travail.

Je remercie Dr Ramzi Farhat, mon encadrant, pour m'avoir proposé le sujet de cette thèse. Je remercie également Dr Maroi Agrebi mon encadrante pour sa patience et son assistance bienveillante.

Je suis honorée de soutenir ce mémoire devant un jury éminent, je vous remercie Madame Marie-Hélène Abel, Madame Lilia Cheniti-Belcadhi, Madame Sonia Ayachi Ghannouchi et Monsieur Faouzi Moussa d'avoir bien accepté d'en faire partie.

A ma famille, toute mon affection.

Dédicaces

« L'éducation est l'arme la plus puissante qu'on puisse utiliser pour changer le monde »

Nelson Mandela

Ce travail s'inscrit dans un processus d'évolution irréversible de l'enseignement. L'avenir appartient à l'école virtuelle.

Cette vision provoque déjà en moi un sentiment de nostalgie, et me rappelle l'ambiance des classes animées, des cours de récréation bruyantes, des concierges, des figures familières de nos maitresses et maitres d'école, professeurs de collège, de lycée et d'université.

Aussi, à la fin de ma scolarité, je leur dédie à toutes et à tous ce mémoire.

A l'âme de ma douce mémé Souad

A la mémoire du bon vivant pépé Mohamed

A ma mémé Jamila, la femme forte et inspirante

et à mon affectueux pépé Mnaouar

Résumé

Dans un contexte de démocratisation du savoir, il est important que la formation en ligne (ou e-learning) se focalise sur le développement de nouvelles approches visant à satisfaire les apprenants tout en minimisant les coûts de formation. De nos jours, le e-learning est confronté à de multiples défis pour répondre aux enjeux actuels qui ont pour ambitions de réduire le taux d'abandon et d'augmenter la satisfaction des utilisateurs. Le développement d'un système qui permet de procurer un contenu éducatif en ligne de qualité est un levier important pour répondre aux besoins des apprenants et des concepteurs pédagogiques. Cette thèse propose un système intelligent d'aide à la décision pédagogique (SIDDP) permettant au concepteur pédagogique d'évaluer le contenu éducatif en ligne dans le but d'améliorer sa production. La problématique majeure pour le développement de ce système est d'objectiver et automatiser la tâche d'évaluation. Notre réponse à cette problématique consiste à concrétiser deux objectifs.

Le premier consiste à proposer une approche d'analyse multicritère des expériences d'apprentissage dénommée MALEA (*Multicriteria Approach for Learning Experience Analysis*). MALEA est adoptée pour l'évaluation de contenus éducatifs en ligne à travers les traces numériques d'interactions des apprenants.

Le second consiste à proposer une approche de prédiction de la réussite des contenus éducatifs en ligne, dénommée ACSP (*Approach for Content Success Prediction*) permettant au concepteur pédagogique d'évaluer son contenu éducatif à n'importe quel stade de son élaboration et notamment avant sa diffusion sur l'Environnement Informatique pour l'Apprentissage Humain (EIAH). Couplant la régression logistique et MALEA, ACSP permet de se prémunir contre l'imprécision éventuelle du jugement humain affectant le processus de décision.

Pour valider expérimentalement l'ensemble de nos approches, deux études de cas ont été effectuées. Une première a été menée sur l'Université Virtuelle de Tunis (UTV). Elle montre, d'une part, que le SIDDP répond à l'objectif recherché et ainsi retenu pour l'évaluation automatique des contenus éducatifs en ligne. D'autre part, elle montre que les résultats obtenus par les tests de performance et l'analyse comparative sont prometteurs avec des valeurs élevées de précision, d'exactitude, de spécificité et de sensibilité. Comme il n'était pas possible de recueillir des données par rapport à la satisfaction des apprenants dans la première étude de cas, une deuxième étude a

été menée dans le but d’expérimenter MALEA avec les quatre critères proposés pour l’analyse des expériences d’apprentissage. Différentes perspectives de recherche sont finalement proposées.
Mots-clés. Analyse de l'expérience d'apprentissage, apprentissage automatique mixte, évaluation de contenu éducatif en ligne, prise de décision pédagogique.

Abstract

E-learning is the future of education. In a context of knowledge democratization, it is important that e-learning focuses on developing new approaches to satisfy learners while minimizing training costs. Nowadays, e-learning is facing huge and multiple challenges and is forced to reduce the dropout rate and increase user satisfaction. Providing high quality online educational contents is an important strategy to support the processes of knowledge and skill acquisition and to meet the expectations of learners and educational designers. Interesting efforts have been made to verify the validity of online educational content, but it is currently a challenging task due to the lack of objectivity and automation in existing researches. In this sense, this study aims at providing an intelligent educational decision support system (IEDSS) allowing educational designers to evaluate their online educational contents in order to improve them.

The major problem for the development of this system is to automate and objectify the evaluation. Our response to this problem consists in concretizing two objectives.

Thus, the first objective of this thesis is to propose a Multicriteria Approach for Learning Experience Analysis (MALEA) on which we have based our online educational content evaluation through the learners' digital traces.

The second objective consists in proposing the Approach for Content Success Prediction (ACSP) that can be used even by educational designers or non-experts to proceed to the automated evaluation of the educational content at any stage of its development and in particular before its diffusion on the Technology Enhanced Learning Environment (TELE). The proposed ACSP approach combines logistic regression and Multicriteria Approach for Learning Experience Analysis (MALEA). This combination helps to guard against the possible imprecision of human judgment affecting the decision-making process.

To experimentally validate our approaches, two case studies were carried out. The first was conducted in the context of the Virtual University of Tunis (VUT). It proves that the IEDSS meets the objective sought and thus retained for online educational content evaluation. On the other hand, results obtained from performance tests and comparative analysis are promising with high values of precision, accuracy, specificity and sensitivity. As it was not possible to collect data about learner satisfaction in the first case study, a second study was conducted to test MALEA with the

four proposed criteria for analyzing learning experiences. Different research perspectives are finally proposed.
Keywords. Learning experience analysis. Blended machine learning. Online educational content evaluation. Educational decision making.

Table de matière

Liste des figures

Liste des tableaux

Liste des abréviations

SIDDP : Système Intelligent d'aide à la Décision Pédagogique

EIAH : Environnement Informatique pour l'Apprentissage Humain

TELE : Technology Enhanced Learning Environment

MALEA : Multi-criteria Approach for Learning Experience Analysis

ACSP : Approach for Content Success Prediction

TIC : Technologies de l'Information et de la Communication

LMS : Système de gestion d'apprentissage (*Learning Management System*)

MOOC : Course en ligne ouvert et massif (*Massive Open Online Course*)

CLOM : Cours en Ligne Ouvert et Massif

SPOC : Small Private Online Courses

COOCs : Corporate Open Online Courses

SNCF : Société nationale des chemins de fer français

LR : régression logistique (*Logistic Regression*)

SVM : machines à vecteurs de support (*Support Vector Machines*)

KNN : k plus proches voisins (*K-nearest neighbors*)

DT: arbre de décision (*Decision Tree*)

RF : forêt aléatoire (*Random Forest*)

DEEDS : *Digital Electronics Education and Design Suite*

ANN : réseaux neuronaux artificiels (*Artificiel Neural Network*)

RNN : réseau de neurones (*Recurrent Neural Network*)

NB : Naïve Bayes

ST : marquage social (*Social Tagging*)

SPM : fouille de motifs séquentiels (*Sequential Patterns Mining*)

DEEDS : *Digital Electronics Education and Design Suite*

SapeS : *Student Academic Performance Evaluation System*

PCA : Analyse en Composantes Principales (*Principal Component Analysis*)

« L'éducation est l'art de faire passer le conscient dans l'inconscient »

Gustave Le Bon

Introduction générale

Contexte et problématique de la thèse

Depuis Caleb Phillips, qui en 1728, proposait dans une annonce parue sur le journal « Boston Gazette », un cours de sténographie par correspondance [Kentnor, 2015], le concept de formation à distance a évolué au grès des progrès techniques. Pendant longtemps, seul le courrier postal assurait la transmission des cours, des devoirs et des corrections. En 1921, les cours étaient diffusés sur les ondes des stations radios. À partir de 1939, ce fut le tour du téléphone et de la télévision. En 1990, avec la popularisation d'Internet et l'avènement du web, la formation en ligne (e-learning) est apparue aux États-Unis et au Canada comme une nouvelle forme de formation.

Le e-learning est médié par des Environnements Informatiques pour l'Apprentissage Humain (EIAH). Ces systèmes informatiques dynamiques et complexes mettent à la disposition des utilisateurs un ensemble de fonctionnalités, d'outils et de ressources dans le but de faciliter la situation d'apprentissage/enseignement. Ainsi, de nombreuses organisations ont mis en place des EIAH visant à ouvrir l'accès à une éducation de qualité tout au long de la vie.

De nos jours, les cours en ligne ouverts et massifs (MOOCs) représentent l'une des formes d'apprentissage en ligne les plus populaires dans le monde, avec un nombre d'inscriptions augmentant exponentiellement. Selon Class Central, ce nombre est passé de 18 millions en 2017 à 180 millions en 2020 [Feng *et al.*, 2019]. En 2021, 40 000 000 nouveaux apprenants se sont inscrits à au moins un MOOC [Class Central, 2022]. Cette demande peut être expliquée entre autres par l'apparition de la pandémie de Covid-19, le e-learning étant devenu incontournable pour les universités, les lycées et les écoles. Les adultes ont également montré davantage d'intérêt à apprendre et partager leurs savoirs, d'une part pour éviter la routine du confinement et d'autre part, pour maintenir le sentiment d'efficacité personnelle et professionnelle. Alors, face à cette croissance spectaculaire de la consommation des formations en ligne, une production massive des contenus éducatifs en ligne a eu lieu.

Malgré ce succès, le taux d'abandon dans la plupart des formations en ligne reste élevé variant d'un EIAH à l'autre. Notamment, environ 7 à 10 % du grand nombre de participants s'inscrivant à des MOOCs parviennent à achever le cours en complétant toutes ses activités [De Freitas *et al.*,

2015] [Gregori *et al*., 2018]. Ce chiffre alarmant a provoqué un débat autour des problèmes du e-learning. De fait, la richesse introduite par les EIAH conduit, dans la pratique, à une certaine complexité. En particulier, avec tous les paramètres qui entrent en vigueur, il est difficile d'identifier et ainsi remédier aux aspects impactant négativement l'expérience d'apprentissage. Des études antérieures ont souligné l'importance d'une conception de qualité des contenus éducatifs en ligne, afin d'améliorer l'apprentissage [Daradoumis *et al*., 2013] [Afify, 2018]. Il manque, cependant, une méthode, un guide, décrivant la façon de les mettre en œuvre [Julia et Marco, 2021].

À l'ère du « *Big Data* », l'accès aux données d'interaction générées au sein des EIAH est devenu une opportunité pour les prestataires de formations en ligne. Deux disciplines issues de la science des données éducatives, émergent : l'analyse de l'apprentissage (*Learning Analytics*) [Gedrimiene *et al*., 2020] et la fouille des données éducatives (*Educational Data Mining*) [Rodrigues *et al*., 2018] [Romero et Ventura, 2020]. L'analyse de l'apprentissage et la fouille des données éducatives s'attachent à dévoiler des informations pertinentes dans le but d'améliorer l'expérience et les environnements d'apprentissage. Ils s'avèrent particulièrement utiles à l'amélioration des contenus éducatifs.

Concrètement, pour améliorer les contenus éducatifs, il faudrait, préalablement, savoir les évaluer [Rodrigues *et al*., 2018]. Jusqu'alors, l'évaluation des contenus éducatifs en ligne, employant ou non les technologies *Learning Analytics* et *Educational Data Mining*, n'a été menée qu'empiriquement à travers des instruments d'enquêtes. Dans certains cas, des questionnaires sont adressés aux apprenants pour enregistrer leur perception. Dans d'autres cas, l'opinion des apprenants n'a pas été prise en compte et en est parfois même exclue. Ces études considèrent que les apprenants n'ont pas l'expertise nécessaire pour évaluer la conception pédagogique [Margaryan *et al*., 2015]. Par conséquent, les chercheurs préfèrent le plus souvent évaluer en s'appuyant sur les connaissances des concepteurs pédagogiques experts observant et interprétant les expériences d'apprentissage des apprenants [Margaryan *et al*., 2015].

Certes, l'analyse empirique est une tâche non-triviale. L'avis des experts ainsi que les perceptions des apprenants sont tous deux utiles pour l'évaluation des contenus éducatifs en ligne. Néanmoins, la situation invite à la réflexion et au questionnement. S'il est permis de douter de la fiabilité de l'avis de l'apprenant, il n'y a pas de raison pour admettre définitivement les résultats d'une expertise, issue d'une observation humaine, probablement incomplète, et une interprétation des analyses empreinte de subjectivité. Ainsi, la complexité de la tâche, associée à la subjectivité humaine, présente un risque

réel pouvant affecter le processus d'évaluation menant à l'amélioration des contenus éducatifs en ligne.

Dans le cadre de cette thèse, c'est dans le contexte d'aide à la décision et d'analyse d'apprentissage que nous intervenons. Nous traitons une problématique sur l'évaluation automatique des contenus éducatifs en ligne dans l'optique d'offrir aux concepteurs pédagogiques une aide pour leurs tâches de conception, d'évaluation et d'amélioration.

Notre objectif principal consiste à proposer un outil permettant d'évaluer automatiquement des contenus éducatifs en ligne. Partant du constat que les méthodes pratiquées d'évaluation, reposant sur le jugement des experts ou des apprenants, présentent un risque de subjectivité, une première question de recherche émerge :

1. Comment évaluer objectivement les contenus éducatifs en ligne ? (QR1)

Nous visons à restreindre la subjectivité des évaluations en cherchant un moyen tenant en compte les expériences d'apprentissage des apprenants à travers leurs traces numériques générées automatiquement dans l'EIAH. A cet égard, nous nous interrogeons :

2. Quels critères peut-on retenir pour une analyse pertinente des expériences d'apprentissage ? (QR2)

Notre objectif est de suivre une démarche d'investigation pour identifier les critères utilisés dans l'état de l'art donnant la possibilité d'analyser et de comprendre les expériences d'apprentissage des apprenants. Nous concluons cette investigation par une synthèse dans laquelle nous ressortons les variables permettant la mesure de chaque critère. Au regard de ce travail de recherche, nous annonçons deux hypothèses :

(H1) : En suivant un contenu éducatif, il y a des apprenants qui génèrent des expériences d'apprentissage positives et ceux qui génèrent des expériences d'apprentissage négatives.

(H2) : Le taux des expériences d'apprentissage positives générées permet d'évaluer le contenu éducatif.

Afin de vérifier ces hypothèses, il nous semble crucial de répondre à cette troisième question de recherche.

3. Comment classifier les expériences d'apprentissage ? (QR3)

Pour répondre à cette question, nous aspirons à recueillir un jeu de données des expériences d'apprentissage et effectuer une classification non supervisée à l'aide d'un algorithme de clustering

performant. Enfin, il nous semble intéressant d'évaluer un contenu éducatif avant de le diffuser sur un EIAH pour améliorer le e-learning. D'où la quatrième question de recherche :

4. Comment prédire le succès/la réussite d'un contenu éducatif en ligne ? (QR4)

En fournissant un ensemble de contenus éducatifs à la machine, nous visons à lui faire apprendre à distinguer un contenu éducatif réussi d'un contenu éducatif à améliorer grâce à un algorithme performant de classification supervisée.

Contributions scientifiques

Afin de faire face à notre problématique, nous proposons un système intelligent d'aide à la décision pédagogique permettant l'évaluation automatique des contenus éducatifs en ligne en vue de recommandations d'amélioration.

- Pour répondre à la première question de recherche, ce système sera fondé sur l'analyse des expériences d'apprentissage. Il combine les deux approches que nous proposons pour répondre à la troisième question et la quatrième question :
- Pour répondre à la deuxième question de recherche, nous identifions les critères d'analyse et les outils de mesure relatives à chaque critère suite à un travail d'investigation et de synthèse fait à partir de l'étude de l'état de l'art.
- Pour répondre à la troisième question de recherche, nous proposons une approche d'analyse multicritère des expériences d'apprentissage (MALEA : *Multicriteria Approach for Learning Experience Analysis*). Celle-ci est basée sur l'algorithme d'apprentissage non supervisé k-means. Cette approche est capable de regrouper les apprenants selon leur comportement. MALEA permet, plus précisément, d'identifier les groupes d'apprenants ayant des expériences d'apprentissage similaires.
- Pour répondre à la quatrième question de recherche, nous proposons une approche de prédiction de la réussite des contenus éducatifs en ligne et de recommandation en vue de leur amélioration (ACSP : *Approach for Content Success Prediction*). Basée sur l'apprentissage automatique supervisé, cette approche permet une classification binaire des contenus éducatifs en ligne. Pour ce faire, nous recourrons à la régression logistique.

Organisation du manuscrit

Composé de quatre chapitres, le mémoire de thèse est organisé comme suit :

1. **L'analyse de l'apprentissage et la fouille des données éducatives au service du e-learning** (chapitre 1). Nous commençons ce chapitre par soulever les défis du e-learning. Nous présentons, ensuite, une étude de la littérature sur la thématique de l'extraction d'information à partir des données éducatives issues des EIAHs. Nous introduisons deux concepts clé du domaine qui sont le *Learning Analytics* et l'*Educational Data Mining*. Nous faisons aussi ressortir la manière dont ces deux approches permettent d'analyser l'apprentissage, dans quels contextes elles s'appliquent et quels sont leurs avantages et leurs limites.
2. **L'aide à la décision dans le domaine du e-learning, vers l'utilisation de l'apprentissage automatique** (chapitre 2). Ce chapitre est dédié à l'état de l'art relatif à l'usage de l'apprentissage automatique (Machine Learning) dans le domaine du e-learning. À travers cette étude, nous exposons les différents paradigmes et techniques de l'apprentissage automatique. Nous mettons l'accent sur les méthodes existantes d'analyse des données éducatives basées sur l'apprentissage automatique. Nous présentons, en outre, les avantages et les inconvénients de certaines méthodes.
3. **Système d'aide à la décision pour l'évaluation et l'amélioration des contenus éducatifs en ligne** (chapitre 3). Après avoir abordé en détail les différents concepts et notions mis en œuvre dans ce travail de recherche, nous réservons le troisième chapitre à la proposition de notre système d'aide à l'évaluation et l'amélioration des contenus éducatifs en ligne. Nous y décrivons son architecture logicielle et ses étapes.
4. **Études expérimentales et analyses des résultats** (chapitre 4). Afin de démontrer sa validation expérimentale et opérationnelle, le système d'aide à la décision proposé a été testé. Les détails et les résultats de l'expérimentation sont présentés et discutés dans ce chapitre.

Nous terminons ce mémoire en fournissant un aperçu global sur les différents travaux réalisés lors de cette thèse avec un regard critique, tout en proposant des perspectives pour des travaux futurs de recherche relatifs à notre problématique.

Chapitre 1 : L'analyse de l'apprentissage et la fouille des données éducatives au service du e-learning

1.1 Introduction

Dans un Environnement Informatique pour l'Apprentissage Humain (EIAH), toute interaction réalisée par un acteur laisse une trace numérique. Lebis regroupe ces traces dans trois catégories. La première correspond aux actions réalisées par les acteurs. La seconde se rapporte aux traces issues de la collecte d'information grâce à un questionnaire soumis aux apprenants. La dernière concerne l'EIAH lui-même, ses ressources et son état [Lebis, 2019]. L'analyse de ces traces aide les décideurs pédagogiques à cerner le comportement des apprenants et ainsi œuvrer à l'élaboration des services améliorant le processus d'apprentissage/enseignement en ligne.

Ce chapitre étudie la capacité de l'analyse des données éducatives à aider dans la prise de décision pédagogique, cadre dans lequel s'inscrivent nos contributions qui se soucient de l'évaluation des contenus éducatifs en ligne à travers l'analyse des expériences d'apprentissage. Ce chapitre introduit, d'abord, dans sa section 2, les notions provenant du domaine du e-learning. Ensuite, la section 3, passe en revue l'utilisation de l'analyse de l'apprentissage (*Learning Analytics*) et la fouille des données éducatives (*Educational Data Mining*) au service du e-learning. Enfin, dans la section 4 et depuis l'état de l'art, sont exposés et discutés les critères et les mesures employés pour analyser les expériences d'apprentissage.

1.2 E-learning : concept et phénomène d'abandon

La diffusion des usages du numérique dans l'enseignement constitue un puissant levier de modernisation, d'innovation pédagogique et de démocratisation du système éducatif. De nos jours, toute personne (élève/étudiant ou autre), quels que soient son âge et son degré d'éducation, a la possibilité d'apprendre et de développer ses connaissances et compétences dans de nombreux domaines. Face à ce changement, le monde se réinvente en permanence et les entreprises ont tout intérêt à intégrer une culture apprenante [Cunningham, 2017].

Offrir des formations en ligne à ses collaborateurs est un investissement durable pour se démarquer de la concurrence. Si le e-learning permet aux salariés d'acquérir du savoir et des aptitudes, afin

de progresser dans leur carrière, il donne également la possibilité aux chercheurs d'emploi de s'adapter aux mutations rapides du marché du travail et de répondre à des défis professionnels réels [Agrebi *et al.*, 2019]. À partir du moment, où elles partent en retraite, de nombreuses personnes trouvent du plaisir à apprendre, jouer d'un instrument, dessiner, faire de la pâtisserie, parler une nouvelle langue, utiliser les nouvelles technologies, etc. Le e-learning peut les encourager à maintenir cet état de curiosité à la nouveauté et au changement tout en offrant un apprentissage en ligne flexible, tout au long de la vie. Ainsi, de nombreuses organisations ont mis en place des EIAH visant à ouvrir l'accès à une éducation de qualité à toute personne. C'est l'ère de la démocratisation du savoir.

1.2.1 Concept du e-learning

- **Définition du e-learning**

Littéralement, le e-learning signifie l'apprentissage / l'enseignement en ligne. Toutefois, il n'est pas aisé de donner une définition exacte et unique du e-learning vue la richesse de ce néologisme. Sangrà *et al.* ont identifié quatre tendances de définitions, chacune repose sur un aspect du e-learning tel qu'illustré dans la Figure 1.1 : la technologie, le savoir, la communication et la pédagogie. La première tendance présente le e-learning comme étant l'utilisation des technologies comme le web et les multimédias dans le but d'apprendre. La seconde tendance décrit le e-learning comme un moyen qui permet l'accès à la connaissance et au savoir. La troisième tendance considère le e-learning comme un outil de communication, d'interaction et de collaboration, tandis que la quatrième tendance définit le e-learning comme une nouvelle approche pédagogique.

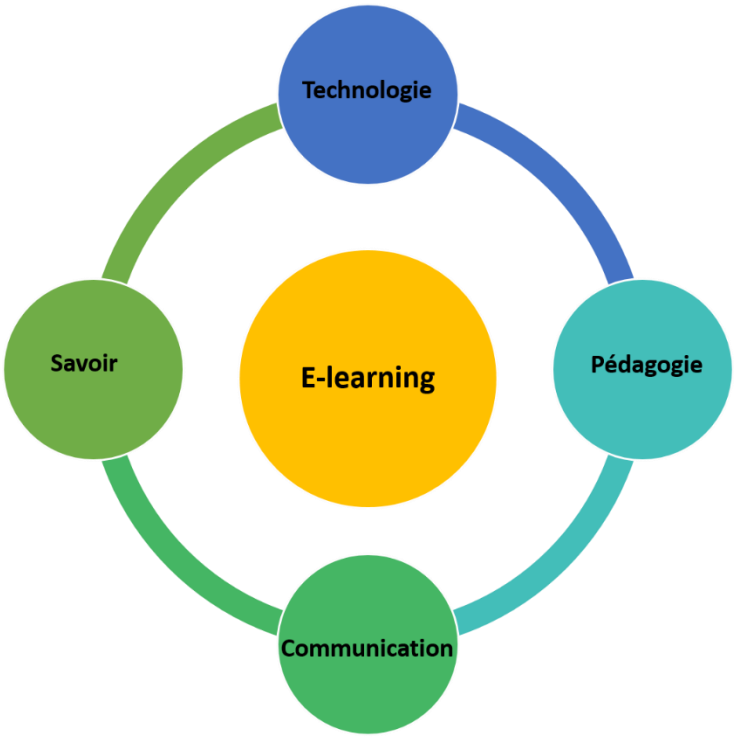

Figure 1.1 Les aspects du e-learning

Cette diversité de définitions a conduit à une variété de dénominations dont la formation à distance, la formation en ligne, la formation ouverte en ligne, la formation ouverte à distance, l'e-formation, l'enseignement à distance, l'enseignement en ligne, etc. [Camargo *et al.*, 2020] [Singh et Thurman, 2019]. Nous employons le terme e-learning dans notre écrit. Par le e-learning, nous désignons l'utilisation des technologies de l'information et de la communication (TIC) pour améliorer la situation de l'apprentissage / l'enseignement en facilitant l'accès à des ressources et des services pédagogiques, ainsi que les échanges et la collaboration en ligne. Nous n'ignorons pas que les besoins d'apprentissage évoluent très rapidement. De ce fait, nous pensons que le concept et les fonctions du e-learning doivent continuellement être adaptés à ces besoins.

- **Environnements Informatiques pour l'Apprentissage Humain (EIAH)**

Afin d'offrir un apprentissage / enseignement en ligne, les institutions mettent en œuvre des EIAH. L'EIAH est un espace virtuel destiné à faciliter l'expérience d'apprentissage / enseignement en ligne [Broisin et Vidal, 2005]. Il fait intervenir essentiellement trois acteurs principaux, que sont respectivement l'apprenant, l'enseignant et l'administrateur [Ouadoud *et al.*, 2021] :

- L'apprenant : consulte en ligne et/ou télécharge les ressources pédagogiques qui lui sont recommandées, organise son travail, effectue des exercices, s'auto-évalue, pose ses questions au tuteur et lui transmet ses travaux.
- L'enseignant : son rôle peut être subdivisé en enseignant concepteur pédagogique, enseignant formateur appelé tuteur, enseignant correcteur, etc. Il crée des parcours pédagogiques, assure le suivi des apprenants et les assiste au cours de l'apprentissage.
- L'administrateur : assure la maintenance du système, s'occupe des inscriptions des apprenants et de la gestion des droits d'accès aussi bien à la plateforme qu'aux ressources pédagogiques.

Les EIAH intègrent des outils pour les différents acteurs du e-learning. L'objectif est de faciliter leurs rôles et fonctions. La Figure 1.2 décrit le modèle d'un EIAH.

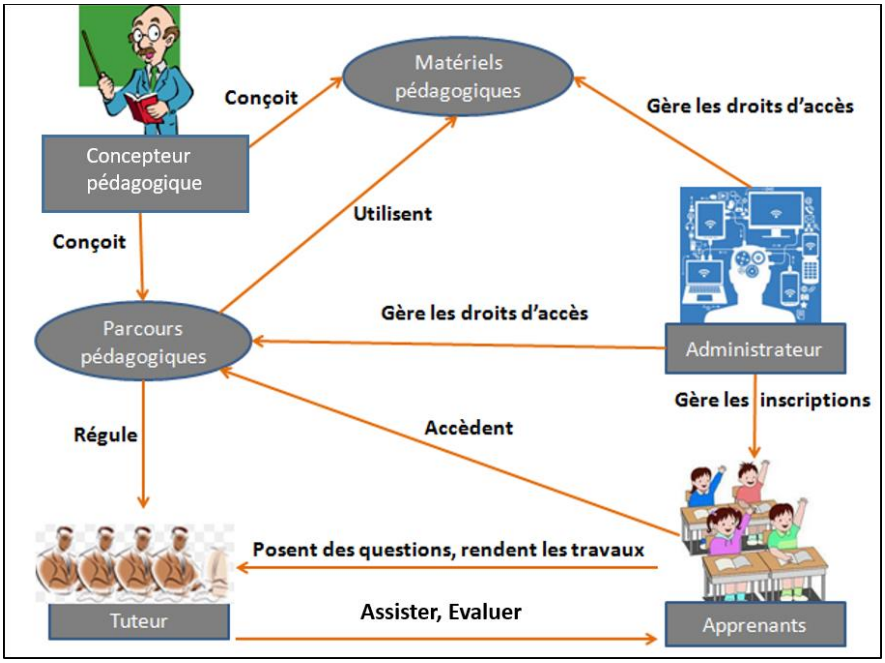

Figure 1.2 Modèle d'un EIAH [George, 2001]

Au début des années 2000, les EIAH sont apparus pour susciter, accompagner et personnaliser l'apprentissage [Balacheff, 2018]. Outre cette fonction d'aide à la réalisation des activités pédagogiques, ils servent également d'outils de présentation de l'information, de communication et de support, aux enseignants et aux apprenants. Dans un EIAH, les modules de cours peuvent

être diffusés via un simple intranet, mais peuvent aussi être gérés par des logiciels spécialisés appelés plateforme du e-learning, système de gestion de formation, système de gestion d'apprentissage (*Learning Management System*), etc. [Alshammari *et al.*, 2018].

Il existe un grand nombre de plateformes de e-learning sur le marché international. Parmi les plateformes sous licence libre, nous pouvons citer Claroline[1], Ganesha[2], et la solution open source Moodle[3], qui a pris son envol dans les années 2000 pour devenir une référence internationale. Il existe également des plateformes propriétaires sous licence comme myTeacher[4], et Blackboard[5]. L'Open University[6] britannique demeure l'un des organismes les plus renommés dans le domaine du e-learning.

- **Cours en ligne ouverts et massifs : MOOCs**

Les *Massive Open Online Courses* (MOOCs) appelés en français Cours en Ligne Ouverts et Massifs (CLOMs) apparaissent à la fin des années 2000. Ils prennent la forme d'EIAH offrant des formations en ligne interactives, ouvertes à tous. Les MOOCs constituent une évolution majeure du e-learning à l'ère d'Internet et représentent l'une des formes de formation les plus populaires dans le monde.

Comme leur nom l'indique, les MOOCs sont ouverts. Ils sont accessibles, gratuitement ou contre paiement, à toute personne disposant d'une connexion Internet et d'un ordinateur, d'une tablette ou bien d'un smartphone [Clarke, 2013]. Depuis 2008, de grandes universités offrent des MOOCs dans un but de démocratisation du savoir. Les apprenants suivent ces cours directement sur les sites de ces établissements, les cas de la FUN[7] (France Université Numérique), de la Sorbonne[8] et de *Harvard*[9]. Les MOOCs peuvent également, être suivis à travers des plateformes dédiées telles que *Udacity*[10], *Coursera*[11], *OpenClassrooms*[12], *edX*[13], etc.

[1] https://www.claroline.com/#/home/accueil
[2] https://www.anema.fr/notre-offre/plateforme-lms/
[3] https://moodle.org/
[4] https://www.myteacher.ch/
[5] https://www.blackboard.com/fr-fr/solutions
[6] https://www.open.ac.uk/
[7] https://www.fun-mooc.fr/fr/
[8] https://www.sorbonne-universite.fr/
[9] https://pll.harvard.edu/
[10] https://www.udacity.com/
[11] https://www.coursera.org/
[12] https://openclassrooms.com/fr/
[13] https://www.edx.org/

La dimension massive est la caractéristique qui distingue le plus le MOOC. Ainsi un MOOC peut accueillir des dizaines de milliers d'apprenants. Par exemple, à la fin de l'année 2017, les plateformes de MOOCs ont proposé 9 400 cours dans le monde entier et attiré 81 000 000 apprenants inscrits en ligne [Feng *et al.*, 2019]. En 2021, 40 000 000 de nouveaux apprenants se sont inscrits à au moins un MOOC [Class Central, 2022]. Par ailleurs, face à l'usage croissant du numérique dans le domaine de l'éducation, certaines/certains enseignantes/enseignants intègrent des MOOCs dans les formations, afin d'expérimenter de nouveaux formats de contenus éducatifs plus interactifs que les vidéos, les documents de type PowerPoint et PDF ou les animations. Alors, de massifs et ouverts, les cours deviennent petits et privés. Nous parlons ici de SPOC (*Small Private Online Courses*), la vitrine des savoirs détenus et diffusés par les enseignantes/enseignants [Kaplan et Haenlein., 2016].

A l'origine, les MOOCs étaient destinés aux étudiants. Aujourd'hui le profil type d'un apprenant MOOC est le salarié qui cherche à développer de nouvelles compétences et valoriser son expérience grâce à un certificat de réussite. D'où l'apparition des COOCs (*Corporate Open Online Courses*). Conçus sur le modèle des MOOCs, les COOCs sont des modules de formation destinés à deux types de public qui sont les salariés d'une entreprise et ses clients. Les COOCs sont de plus en plus utilisés dans les grandes entreprises comme Renault et la SNCF pour former les salariés et fidéliser la clientèle [Miss MOOC, 2016] [Renault Groupe, 2018].

1.2.2 Phénomène d'abandon

- **Présentation**

Malgré le succès du e-learning s'est accru considérablement pendant la pandémie de Covid-19, les individus engagés, qui continuent jusqu'au bout la formation, sont loin d'être majoritaires. Notamment, environ 7 à 10 % du grand nombre de participants s'inscrivant à des MOOCs parviennent à achever les cours en complétant toutes les activités [De Freitas *et al.*, 2015 ; Gregori *et al.*, 2018 ; Aldowah *et al.*, 2020]. Ainsi, une attention particulière est accordée au phénomène de l'abandon appelé aussi non-persistance. Le chiffre est explicable puisque les motivations de suivre un cours ou une formation en ligne sont diverses et multiples, allant de la simple curiosité pour la thématique générale d'un cours, à l'envie d'acquérir des connaissances et des compétences sans être engagé dans un rythme de travail régulier.

L'abandon scolaire, terme utilisé à la fois pour le secondaire, le collégial et l'universitaire est communément perçu comme étant l'interruption temporaire ou définitive des études avant l'obtention d'une reconnaissance des acquis, un diplôme, un certificat, une attestation d'études ou autre, de la part d'une institution d'enseignement confirmant la fin des études [Tsolou *et al*., 2021]. Deux types d'abandon sont constatés. La non-participation suite à l'inscription au cours et le retrait volontaire hâtif ou à un stade quelque peu avancé [Alamri *et al*., 2019].

- **Facteurs d'abandon**

Les écrits scientifiques évoquant les facteurs qui exercent une influence sur le comportement de l'apprenant dans les activités d'apprentissage en ligne provoquant particulièrement l'abandon utilisent le plus souvent les termes de barrière, facteur et obstacle. Dans les travaux doctoraux de Cisel, l'auteur dénombre cinq barrières potentielles : la barrière situationnelle, la barrière dispositionnelle, la barrière institutionnelle, la barrière épistémique et la barrière technologique [Cisel, 2016].

Les barrières d'ordre situationnel sont liées aux circonstances particulières de la vie de l'apprenant, notamment du fait de responsabilités familiales, d'engagement professionnel ou social ou de problèmes de santé. Par conséquence, le manque de temps peut conduire à un retrait volontaire ou à une non-participation.

Les barrières d'ordre dispositionnel appelées aussi barrières psychosociales, renvoient aux croyances, perceptions, valeurs, et attitudes qu'entretient une personne en regard de sa participation aux activités de formation. Une personne peut par exemple interrompre sa formation, car elle se considère, à un moment donné, trop âgée pour la poursuivre ou par la crainte d'échouer. La perception négative de soi, découlant généralement des expériences scolaires antérieures, créent aussi une entrave [Mystakidis *et al*., 2021].

Dans le cas du e-learning, l'apprenant se trouve exposé à une double contrainte, celle de devoir intégrer des connaissances nouvelles sur un domaine donné et en même temps celle d'utiliser un dispositif technique pour ce faire. Cette tâche est complexe si l'EIAH n'est pas facile d'utilisation par exemple. Nous parlons bien des barrières technologiques.

Quant aux barrières d'ordre institutionnel, elles dépendent de l'institution et des caractéristiques de la formation comme le contenu et le rythme du cours, la qualité des ressources pédagogiques,

la réactivité de l'équipe pédagogique, le mode d'apprentissage et même les modalités d'évaluation puisque certains apprenants se retirent de la formation au moment de l'évaluation.
La motivation et les motifs d'entrée se révèlent, en outre, comme des facteurs déterminants, voire décisifs. D'après l'étude de Dogbe-Semanou portant sur des apprenants à distance en Afrique subsaharienne francophone, Dogbe-Semanou affirme que la première des choses qui expliquent la persévérance ou l'abandon au Togo est la motivation liée à la possibilité d'utiliser, à des fins professionnelles, les compétences acquises [Dogbe-Semanou, 2016].
Les facteurs d'abandon étant nombreux et variés, nous nous proposons de les classifier, sommairement selon deux catégories, en facteurs liés aux apprenants d'une part, en facteurs liés aux EIAH d'autre part [Mourali *et al*., 2020a]. Dans ce mémoire, nous nous intéressons aux facteurs liés aux EIAH et plus précisément aux problèmes en relation avec les contenus éducatifs en ligne.

1.3 Analyse des données éducatives : défis et enjeux

Avec la digitalisation de la société, et grâce au développement de la capacité de stockage et de la puissance de calcul, tout ce que nous utilisons produit des données : téléphones, objets connectés, capteurs, cartes bancaires, cartes de fidélité, sites web, applications, etc. Ces données générées, décrivant nos activités et fournissant une multitude d'informations, valent de l'or. D'ailleurs, la métaphore « pétrole du XXI^{e} siècle » s'est imposée dans la littérature [Lévy, 2021] et les données finissent par être considérées comme le miroir de l'économie à l'échelle mondiale quels que soient les secteurs. Cette nouvelle ressource, en pleine expansion, demeure un fondement de création de richesse et par conséquent un facteur déterminant de la compétitivité des entreprises et des Etats. La gratification de la richesse ne se trouve pas dans la simple possession des données. Tout l'enjeu réside dans la capacité à donner du sens à cette mine d'informations.
A ce stade, une question se pose, notamment : quel bénéfice peut-on tirer des données ? Nous citons ici des exemples concrets des bénéfices des données dans des domaines variés. A titre d'exemple l'exploitation des données médicales présente de nombreux intérêts dont l'identification de facteurs de risque de développement de maladie qui permet de mettre en place des outils de prévention. L'accès aux données peut contribuer au développement de nouveaux produits de santé comme il peut être une aide au diagnostic, au choix et au suivi de l'efficacité des traitements. Les décideurs du secteur se servent aussi des données médico-administratives pour réduire les dépenses de la santé publique [Srinivasan et Arunasalam, 2013] [Dimitrov, 2016] [Dash

et al., 2019]. Dans le domaine du e-commerce, il est crucial d'analyser les comportements des consommateurs pour pouvoir proposer des offres personnalisées et des services optimaux [Akter *et al.*, 2016] [Chen, 2018] [Malhotra et Rishi, 2021]. La bonne connaissance des clients permet, aussi, de fournir aux banques et aux compagnies d'assurance des indicateurs précieux pour minimiser leurs risques financiers [Hussain et Prieto, 2016]. Même en politique, grâce aux données personnelles collectées sur les votants, un candidat aux élections peut cibler aux mieux ses électeurs clés [Sudhahar *et al.*, 2015] [Ceron et Iacus, 2016] [González, 2017]. Comme tous les domaines, l'éducation n'échappe pas à l'abondance et la prolifération de données. Les EIAH eux aussi produisent des quantités énormes de données sur les contenus éducatifs en ligne, les apprenants et leurs interactions [Adam *et al.*, 2018] [Popchev et Orozova, 2019]. Nous souhaitons, dans cette étude, découvrir la valeur cachée des données éducatives et contribuer à soutenir le e-learning.

1.3.1 La fouille des données éducatives et l'analyse de l'apprentissage

Lorsque les apprenants accèdent aux EIAH, ils laissent automatiquement des traces numériques enregistrées, constituant des données qui décrivent leurs activités, leur comportement et bien d'autres informations. L'intelligence artificielle a le pouvoir de conférer un sens à toutes ces données. En éducation, l'intelligence artificielle se manifeste sur deux champs : l'analyse de l'apprentissage (*Learning Analytics*) et la fouille de données éducatives (*Educational Data Mining*) [Lemay *et al.*, 2021].

L'*Educational Data Mining* s'intéresse à l'application et au développement des techniques de la fouille de données (*Data Mining*) pour explorer les données éducatives [Romero et Ventura, 2020]. L'*Educational Data Mining* se préoccupe particulièrement, de la recherche des modèles et des associations de données permettant de tirer des conclusions sur le comportement des apprenants et la performance des EIAH. Le *Learning Analytics* est un champ émergent à la confluence de plusieurs disciplines telles que les statistiques, le *Data Mining*, la visualisation de données, l'apprentissage automatique, la psychologie, l'analyse des réseaux sociaux, etc. [Avella *et al.*, 2016]. « Il traite la mesure, la collecte, l'analyse et la communication de données sur les apprenants et leurs contextes, dans le but de comprendre et optimiser l'apprentissage et les environnements dans lesquels il est produit » [Gedrimiene *et al.*, 2020].

Bien que chacun de ces deux champs ait sa spécificité, *l'Educational Data Mining* et le *Learning Analytics* se partagent cependant certains objectifs qui les conduisent à collaborer et à s'enrichir mutuellement. En effet, tous deux visent à rechercher des informations pertinentes dans les données éducatives, afin d'étayer et d'orienter les décisions liées au secteur de l'éducation [Almohammadi *et al.*, 2017]. Les applications de l'*Educational Data Mining* et du *Learning Analytics* dans le domaine du e-learning sont nombreuses et variées. La Figure 1.3 les classe selon trois catégories : analyse des sentiments, apprentissage autorégulé et prédiction du comportement de l'apprenant.

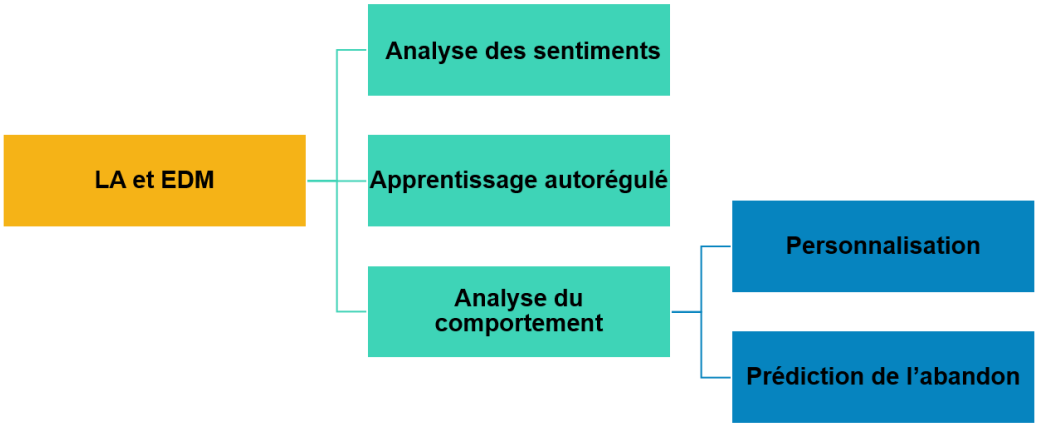

Figure 1.3 Application de l'Educational Data Mining et du Learning Analytics dans le domaine du e-learning

1.3.1.1 Analyse des sentiments

L'analyse de l'attitude des apprenants, l'identification de leur perception et même de leurs émotions à l'égard de l'apprentissage, contribuent à l'amélioration du processus pédagogique dans le contexte du e-learning [Georgescu et Bogoslov, 2019].

Des travaux de recherche se sont appuyés sur la mesure du niveau d'engagement des apprenants pour mettre à jour et enrichir le contenu éducatif. Moubayed *et al.* utilisent des algorithmes d'apprentissage automatique non supervisé pour identifier le niveau d'engagement des apprenants [Moubayed *et al.*, 2020]. Ils implémentent k-means [14] pour regrouper les apprenants sur la base de leurs activités et interactions en ligne. Les résultats montrent que, si le modèle à deux niveaux

[14] *K-means* ou k-moyenne est un algorithme d'apprentissage automatique qui permet de regrouper les données dans des groupes similaires appelés clusters. Nous donnons plus de détails sur cet algorithme aux chapitres 3 (sections 3.3.2 et 3.3.3) et 4 (section 4.2.1).

d'engagement est le plus performant en termes de séparation des groupes, le modèle à trois niveaux d'engagement est le meilleur modèle à adopter pour les instructeurs afin d'identifier les apprenants non engagés. L'une des perspectives de ce travail de recherche est de déterminer les composantes les plus ou les moins engageantes dans un cours en ligne.

Awidi *et al*. intègrent les étudiants dans un groupe Facebook afin d'accentuer leur engagement dans un cours d'architecture initié par l'Université d'Australie occidentale [Awidi *et al*., 2019]. Ils mènent ensuite une enquête pour appréhender la perception des étudiants sur leurs expériences d'apprentissage suite à l'intégration du groupe Facebook. Parmi un nombre total de 108 étudiants inscrits au cours, 60 étudiants ont répondu aux questions de l'enquête. La majorité a donné des réponses de satisfaction favorables au cours remanié. La limite de cette étude réside dans les données utilisées. En effet, les réponses obtenues peuvent ne pas refléter fidèlement les véritables perceptions puisqu'un nombre important d'étudiants ne se sont pas exprimés.

Krithika et Lakshmi Priya proposent un système capable d'analyser l'excitation et la perturbation d'un apprenant suivant un cours en ligne, à travers le mouvement de la tête et des yeux [Krithika et Lakshmi Priya, 2016]. Ce système permet de classifier les apprenants selon l'implication et l'intérêt au sujet du cours. Les résultats de cette classification sont fournis, en temps réel, comme un retour (*feedback*) pour assurer une meilleure livraison de contenus éducatifs. A l'effet d'étudier l'apport de ce système, un questionnaire de satisfaction comportant 132 questions a été adressé aux apprenants. Indépendamment de l'aspect subjectif des réponses, puisqu'il s'agit de perception humaine, le nombre d'items est assez élevé ce qui pourrait affecter les données mises à disposition de l'étude.

Hew *et al*. se sont intéressés à la mesure du succès des MOOCs à partir de l'analyse des perceptions des étudiants [Hew *et al*., 2021]. À cette fin, ils ont adopté un algorithme d'apprentissage automatique supervisé, une analyse des sentiments et une modélisation linéaire hiérarchique. Il semble que les apprenants qui ont abandonné leurs cours étaient moins susceptibles de fournir des commentaires sur leurs expériences d'apprentissage. Par conséquent, nous pensons que la prédiction du succès des MOOCs nécessite d'autres éléments outre la satisfaction des apprenants.

1.3.1.2 Apprentissage autorégulé

L'analyse des données éducatives est pareillement utilisée pour améliorer l'apprentissage autorégulé. L'apprentissage autorégulé consiste à prendre conscience de la manière d'apprendre,

se fixer des objectifs pédagogiques et sélectionner des stratégies pour les atteindre sans être guidé par un tuteur [Wong *et al*., 2019]. Comme rapporté par Kivimäki *et al*., plusieurs EIAH proposent aux apprenants des outils de mesure, afin de les aider à planifier et organiser leurs apprentissages, à faire des auto-évaluations et à produire des analyses personnalisées sur leurs activités d'apprentissage [Kivimäki *et al*., 2017]. Dans cette direction, face aux faibles performances des étudiants de première année inscrits aux cours d'apprentissage des langues assisté par ordinateur (*Computer-Assisted Language Learning*) à l'Université de Kyushu au Japon, Li *et al*. proposent un système d'apprentissage autorégulé basé sur le *Learning Analytics* [Li *et al*., 2017]. Ce système est composé de deux parties. La première partie est dédiée aux apprenants. Elle leur permet de visualiser des aspects de leurs comportements et d'effectuer un autocontrôle. Les apprenants peuvent ainsi avoir des mesures sur la progression de leurs apprentissages. La seconde partie est destinée aux instructeurs. Elle leur permet d'identifier les modèles de comportement d'apprentissage et de prendre des décisions éclairées.

Dans un cours hybride destiné aux enseignants de musique, Montgomery *et al*., recueillent des données sur les accès et le temps passé sur la plateforme Moodle [Montgomery *et al*., 2019]. L'objectif est de comprendre le comportement d'autorégulation que les apprenants utilisent pour soutenir leurs apprentissages en ligne. Une corrélation significative entre le comportement d'autorégulation et la réussite scolaire est détectée. Cependant, les chercheurs ne sont pas arrivés à expliquer pourquoi les apprenants régulent leurs apprentissages d'une certaine manière plutôt que d'une autre. Ceci revient au pouvoir limité des données explorées dans cette étude de cas. Par conséquent, les instructeurs ne seront pas en mesure d'intervenir pour améliorer le cours conformément aux besoins des apprenants [Montgomery *et al*., 2019].

Kizilcec *et al*. ont procédé à une investigation sur l'apprentissage autorégulé à travers six MOOCs en se basant sur la réussite du cours, les interactions avec le contenu du cours et les réponses des apprenants à une enquête [Kizilcec *et al*., 2017]. Cette étude a montré que les données démographiques ainsi que la motivation des apprenants, permettent de prédire les compétences et les capacités des apprenants en matière d'autorégulation d'apprentissage.

En conclusion, la valorisation des données éducatives dans le contexte de l'apprentissage autorégulé est une tendance. Les recherches dans le champ du *Learning Analytics* ont été menées principalement pour mesurer plutôt que pour soutenir l'apprentissage autorégulé. Il existe donc un

besoin crucial d'exploiter le *Learning Analytics* pour favoriser l'apprentissage autorégulé des apprenants dans les EIAH [Viberg *et al*., 2020].

1.3.1.3 Analyse du comportement

L'analyse du comportement des utilisateurs en ligne est un domaine de recherche en pleine expansion. Elle s'avère utile à la prise des décisions éclairées dans la stratégie pédagogique [Lung-Guang, 2019].

Ainsi, en combinant à la fois les données hors ligne des apprenants y compris les caractéristiques et les données démographiques, et les données en ligne comme les journaux d'activités, Azcona *et al*. construisent un modèle prédictif capable de distinguer les apprenants en besoin d'aide lors de l'apprentissage de ceux qui risquent d'échouer à leur prochaine évaluation [Azcona *et al*., 2019]. Plusieurs algorithmes de classification ont été appliqués comme la régression logistique (*Logistic Regression*) [Catal *et al*., 2019] [Kohli *et al*., 2021], les machines à vecteurs de support (*Support Vector Machines*) [Hasan et Boris, 2006] [Chui *et al*., 2020], la méthode des k plus proches voisins (*K-nearest neighbors*) [Sathish *et al*., 2020] [Taunk *et al*., 2019], l'arbre de décision (*Decision Tree*) [Berry *et al*., 2019] [Charbuty et Abdulazeez, 2021] et les forêts aléatoires (*Random Forest*) [Biau et Scornet, 2016]. Le classificateur à 12 plus proches voisins donne les meilleures performances en termes de précision et de rappel. La solution proposée pour soutenir les apprenants dans cette étude consiste à leur envoyer des notifications personnalisées. Cependant, étant donné le nombre très élevé d'e-mails que les apprenants reçoivent chaque jour de la part des enseignants, de l'association des étudiants, de l'administration, etc., ces auteurs mettent en avant le fait que certains apprenants n'arriveront peut-être jamais à regarder la totalité de ces notifications.

En utilisant les données enregistrées par un EIAH appelé *Digital Electronics Education and Design Suite* (DEEDS), Hussain *et al*. visent à prédire les performances des étudiants ainsi que les difficultés qu'ils peuvent rencontrer dans une session de cours de conception numérique [Hussain *et al*., 2019]. Pour ce faire, cinq algorithmes d'apprentissage automatique ont été implémentés et comparés : réseaux neuronaux artificiels (*Artificiel Neural Network*), machines à vecteurs de support (*Support Vector Machines*), régression logistique (*Logistic Regression*), Naïve Bayes et arbre de décision (*Decision Tree*). Les ANN et les SVM obtiennent la précision la plus élevée (75 %). La limite principale de cette étude est liée aux caractéristiques utilisées pour la classification.

Ces caractéristiques sont très spécifiques au cours sélectionné. Par conséquent, le modèle de classification proposé n'est pas généralisable.

Alharbi *et al*. visent à identifier, le plus en amont possible, les apprenants qui risquent d'obtenir de mauvaises notes aux examens [Alharbi *et al*., 2016]. Aussi, ils recourent à des techniques d'*Educational Data Mining* à travers l'outil IBM SPSS Modeler v15 permettant de combiner plusieurs modèles de classification comme la régression logistique, le réseau de neurones, et les quatre arbres de décision C5, C&R, Quest et CHAID, afin d'identifier les apprenants à risque de performance académique faible. Les données recueillies pour cette étude sont les données à l'admission ainsi que les résultats des modules de la première année. Dans leurs perspectives, les chercheurs envisagent l'intégration des données relatives à l'engagement des apprenants afin de vérifier leur impact sur la précision des prédictions.

Mubarak *et al*. ont employé les réseaux de neurones profonds pour analyser les données du flux de clics vidéo [Mubarak et Ahmed, 2021]. Le but de cette étude consiste à prédire les performances hebdomadaires des apprenants pour permettre aux instructeurs la possibilité d'une intervention opportune.

Outre les exemples précédents, l'analyse du comportement des apprenants a été largement utilisée dans la personnalisation de l'apprentissage ainsi que dans la prédiction de l'abandon.

- **Personnalisation**

L'intérêt pour le *Learning Analytics* et l'*Educational Data Mining* est suscité notamment par le besoin de personnalisation et d'adaptation des contenus éducatifs en fonction des profils, préférences et niveau de connaissances des apprenants, éléments discernables grâce à l'analyse du comportement [Klašnja-Milićević *et al*., 2017] [Agrebi *et al*., 2019].

Des études ont été menées pour faire évoluer les systèmes de recommandation. Klašnja-Milićević *et al*. proposent une méthode hybride combinant le marquage social (*Social Tagging*) et la fouille de motifs séquentiels (*Sequential Patterns Mining*) [Klašnja-Milićević *et al*., 2018] afin de recommander les ressources pédagogiques les plus adaptées aux apprenants. Cette approche est coûteuse en temps et fastidieuse car la catégorisation des tags n'est pas automatisée.

Belarbi *et al*. proposent un système de recommandation de vidéos à travers un petit cours privé en ligne (SPOC) [Belarbi *et al*., 2018]. A cette fin, ils analysent, d'abord, le comportement des apprenants lors du visionnage des vidéos pour estimer leur intérêt. Ils identifient ensuite, les

apprenants ayant des profils similaires en utilisant l'algorithme de *clustering* k-means. Enfin, ils recommandent à l'apprenant cible les mêmes vidéos, auxquelles les apprenants similaires ont manifesté de l'intérêt. Dans cette étude de cas, l'évaluation de la pertinence des recommandations est absente.

Gonçalves *et al*. offrent une solution pour améliorer les performances des apprenants lors des tests [Gonçalves *et al*., 2018]. Ils proposent une nouvelle approche soutenue par le *Learning Analytics* et un système de recommandation basé sur les ontologies pour personnaliser les évaluations. Dans la première phase de l'approche, les chercheurs analysent les réponses des apprenants aux tests. Dans la deuxième phase, une base de connaissances est établie grâce à l'ontologie alimentée par les réponses des apprenants. Cette base de connaissance sert à recommander aux apprenants de vérifier leurs réponses quand elles sont incorrectes. De cette manière, l'apprenant dispose d'une nouvelle tentative avant de confirmer définitivement la réponse à une question spécifique. Les réponses originales à chaque question et la réponse obtenue après une recommandation donnée peuvent fournir des informations intéressantes pour les tuteurs afin de cerner les difficultés rencontrées et formuler des stratégies supportant l'apprentissage.

Pour aider les tuteurs dans le processus d'évaluation, Costa *et al*. ont mis en œuvre une architecture logicielle appelée *Student Academic Performance Evaluation System* (SapeS) fondée sur le *Learning Analytics* et les objectifs d'apprentissage [Costa *et al*., 2019]. En utilisant les interactions des apprenants, SapeS permet aux tuteurs de suivre la progression des apprenants et de réadapter le plan pédagogique en fonction du niveau de performance de chacun. Cet outil doit être validé expérimentalement pour s'assurer de sa contribution effective dans l'aide à l'évaluation et au soutien des performances académiques.

Bourkoukou et El Bachari proposent le système d'apprentissage adaptatif *LearnFitII (framework for adaptive learning management system)* [Bourkoukou et El Bachari, 2018]. Celui-ci vise à reconnaître le profil de l'apprenant et de recommander un ensemble approprié d'objets d'apprentissage pour améliorer le processus d'apprentissage. A cet effet, *LearnFitII* propose d'abord, un scénario d'apprentissage personnalisé pour retenir les nouveaux apprenants. Chaque apprenant qui accède à *LearnFitII* pour la première fois est invité à passer un test de style d'apprentissage (questionnaire psychologique) basé sur l'approche de Felder et Silverman [Felder et Silverman, 1988] afin d'identifier son style d'apprentissage et ses habitudes dépeignant son profil. Le style d'apprentissage d'un individu étant la manière dont cette personne est programmée

pour apprendre le plus efficacement, c'est-à-dire pour recevoir, comprendre, retenir et être capable d'utiliser une nouvelle information [Reinert, 1976]. Les chercheurs explorent, ensuite, les données de journaux à l'aide de l'algorithme k plus proches voisins (KNN) et la fouille des règles d'association pour recommander les objets d'apprentissages.

Selon [Kaabi *et al*., 2020], le manque d'interactivité personnalisée dans les EIAH peut causer un taux d'abandon élevé. Dans le dessein de remédier à cette situation, les chercheurs se sont engagés à personnaliser l'apprentissage. Ils déterminent dans un premier les profils des apprenants à travers la version arabe du questionnaire FSLSM (*Felder-Silverman learning style model*). Tenant compte des informations sur chaque profil, Kaabi *et al*. proposent, dans un deuxième temps, une approche soutenant la décision de l'enseignant sur la stratégie de personnalisation appropriée à considérer pour un cours donné. Cette approche permet d'attribuer automatiquement, à chaque apprenant, de nouvelles activités et de nouveaux parcours. A cet égard, une classification non supervisée basée sur l'algorithme k-means est utilisée pour regrouper les apprenants selon les similitudes et les différences en fonction du profil. Il était souhaitable de tester d'autres algorithmes de classification non supervisée (*clustering*) et de comparer leurs performances.

- **Prédiction de l'abandon**

Outre la personnalisation, le *Learning Analytics* et l*'Educational Data Mining* ont été employés pour prédire les cas d'abandon afin d'accrocher davantage les apprenants. Selon Foster et Siddle, le *Learning Analytics* permet d'identifier les apprenants à risque de quitter la formation en ligne et s'avère plus efficace que le ciblage basé sur des données démographiques [Foster et Siddle, 2020]. Sur la base des notes d'évaluation, Chui *et al*. recourent aux machines à vecteurs de support (SVM) pour prédire les apprenants marginaux ceux susceptibles d'échouer et ceux à risque certain d'échouer à travers leurs notes d'évaluation [Chui *et al*., 2020]. Les résultats montrent que la méthode proposée est capable d'atteindre une précision globale de 92,2-93,8% pour la prédiction des apprenants à risque certain d'échouer et de 91,3-93,5% pour la prédiction des apprenants marginaux. Les chercheurs soulignent qu'il est possible d'améliorer la précision de la classification, mais ne donnent pas plus de précision à ce sujet. Ils suggèrent aussi de tester cette méthode dans d'autres applications de *Learning Analytics*.

Timbal présente un modèle prédictif intelligent basé sur une méthode de classification supervisée tel que l'arbre de décision (DT) pour prévoir qui, parmi les apprenants inscrits, risque de décrocher

[Timbal, 2019]. Dans cette étude de cas, les auteurs explorent l'ensemble de données SARDO (*Student At-Risk of Dropping Out*) relatif aux élèves de terminale de l'école secondaire nationale de Kapalong de la province philippine Davao du Nord, ayant abandonné leurs études. Des données relatives aux résultats scolaires et aux aspects démographiques ont été explorées. L'évaluation de la performance du modèle prédictif intelligent proposé est souhaitée.

1.3.2 Critères et mesures d'analyse des expériences d'apprentissage

La notion d'expérience d'apprentissage désigne la conscience préréflexive qui accompagne l'activité d'apprentissage, c'est-à-dire l'état psychique de l'apprenant dans lequel il a conscience de son statut d'apprenant et du processus d'apprentissage en même temps. Elle comporte des dimensions intentionnelles, perceptives, émotionnelles et cognitives [Dieumegard et Durand, 2005]. Nous nous investissons dans l'analyse des expériences d'apprentissage car susceptible d'apporter une aide à la prise de décision pédagogique.

Nous réalisons, après le passage en revue de l'état de l'art en matière d'analyse de données éducatives, qu'il est judicieux d'asseoir l'analyse des expériences d'apprentissage sur un faisceau de quatre critères : engagement de l'apprenant, satisfaction de l'apprenant, évaluation des performances de l'apprenant aux tests et examens et achèvement de la formation en ligne.

En effet, notre étude de l'état relève ces quatre critères mesurables employés dans l'analyse des expériences d'apprentissage. La synthèse qui nous a permis d'identifier ces quatre critères est détaillée dans la section 1.3.3.

L'engagement, premier critère suggéré, est défini comme étant le degré d'intérêt que l'apprenant manifeste pour la formation en ligne [Moubayed *et al*., 2020]. Selon Kori *et al*., il a un impact significatif sur les décisions des apprenants [Kori *et al*., 2016]. L'engagement peut être observé à travers d'une part le degré d'interaction de l'apprenant avec le contenu éducatif, et d'autre part l'effort fourni pour accomplir les tâches connexes [Moubayed *et al*., 2020]. Dans cette définition, Moubayed *et al*. insistent sur l'engagement comportemental et cognitif car ils fournissent une vision quantifiable de l'engagement. Ainsi, pour quantifier l'engagement, Moubayed *et al*. proposent des mesures liées à l'interaction et des mesures liées à l'effort. Les mesures liées à l'interaction décrivent la fréquence à laquelle l'apprenant a interagi avec le contenu éducatif diffusé sur l'EIAH, tandis que les mesures liées à l'effort décrivent le volume d'effort fourni par l'apprenant pour accomplir les activités d'apprentissage.

Le second critère retenu est la satisfaction l'émotion ayant une implication sur les dispositions de l'apprenant au cours du processus d'apprentissage [Imani et Montazer, 2019]. Selon McGillicuddy, il est crucial que l'apprenant développe une attitude positive envers son expérience d'apprentissage [McGillicuddy, 2020]. Pham *et al.* montrent à travers une enquête auprès de 1232 étudiants vietnamiens que la satisfaction dépend de la qualité des services offerts par les EIAH. Un enjeu primordial du e-learning consiste à motiver les apprenants afin de leur procurer de la satisfaction et ainsi les fidéliser [Pham *et al.*, 2019].

Le troisième critère relevé est celui de l'évaluation de la performance des apprenants aux tests et examens. Son utilisation a concouru par exemple pour s'enquérir de l'efficacité du e-learning adaptatif [Hubalovsky *et al.*, 2019], et accompagner les apprenants dans leur progression [Alharbi *et al.*, 2016] [Chui *et al.*, 2020] [Almaiah et Alyoussef, 2019], etc.

Enfin, un quatrième critère d'analyse des expériences d'apprentissage, l'achèvement de la formation en ligne. Il a été exploité dans le but d'approfondir la compréhension et ainsi la prévention des facteurs d'abandon [Dalipi *et al.*, 2018] [Kaabi *et al.*, 2020] [Foster et Siddle, 2020] [Timbal, 2019].

Nous nous attelons à relever les critères mesurables d'analyse des expériences d'apprentissage utilisés dans la littérature et les exposer dans le Tableau *1.1* ci-dessous.

Tableau 1.1 Critères et mesures utilisés dans la littérature pour analyser les expériences d'apprentissage en ligne.

Critères d'analyse des expériences d'apprentissage	**Mesures**
Engagement	- Mesures liées aux interactions : Nombre de questions posées [Reid, 2012], Nombre de participations [Reid, 2012], Nombre d'interactions avec le tuteur [Reid, 2012], Nombre de publications dans les forums [Ramesh *et al.*, 2013], Indicateur binaire de l'achèvement de l'examen [Ramesh *et al.*, 2013], Nombre d'accès au cours [Moubayed *et al.*, 2020], Nombre de ressources téléchargées [Moubayed *et al.*, 2020], Nombre de publications lus [Moubayed *et al.*, 2020]. - Mesures liées à l'effort fourni : Temps écoulé dans le cours [Reid, 2012],

	Nombre de consultations du contenu [Ramesh *et al.*, 2013], Temps écoulé dans l'EIAH [Kim *et al.*, 2016], Nombre de visites de l'EIAH [Kim *et al.*, 2016], Indicateur binaire de soumission de devoir à temps [Moubayed *et al.*, 2020], Durée moyenne en heures entre l'affichage et la remise des devoirs [Moubayed *et al.*, 2020].
Satisfaction	Avis des apprenants observés à travers : - L'évaluation par étoiles avec échelle de 2 à 10 - L'échelle d'évaluation de la satisfaction avec smiley - Les questionnaires - Les enquêtes
Evaluation de la performance	Notes aux examens [Herodotou *et al.*, 2019] Scores dans les quiz Indicateur binaire du résultat de l'examen final (échec/ réussite)
Achèvement de la formation	Indicateur binaire de l'achèvement de l'examen final [Ramesh *et al.*, 2013] Progression de l'apprentissage en termes de pourcentage

D'après ce tableau, nous avons pu identifier, à travers l'étude de l'art, les mesures possibles permettant d'évaluer les expériences d'apprentissage des apprenants. Certaines mesures ont été employées pour l'évaluation d'un ou de plusieurs critères. Mais dans le cadre de notre étude nous ne nous intéressons pas à ce détail car nous proposons une évaluation multicritère. Ainsi, toute mesure est la bienvenue, peu importe son attribution à un critère ou à un autre. Nous considérons dans notre réponse à la première question de recherche que l'expérience d'apprentissage dans sa globalité permet d'évaluer objectivement le contenu éducatif en ligne.

1.3.3 Discussion

Le *Learning Analytics* et l'*Educational Data Mining* ont contribué substantiellement au développement du e-learning. Comme précisé tout au long de ce chapitre, les travaux de recherche ont principalement exploré les données des apprenants pour analyser, comprendre et évaluer les expériences d'apprentissage. Dans ce sens de nombreuses approches ont été adoptées : des approches basées sur l'observation et le jugement humain [Awidi *et al.*, 2019] [Margaryan *et al.*,

2015], d'autres sur les statistiques [Aboagye *et al*., 2020] [Lara et Pamplona, 2020], l'exploration de données [Salloum *et al*., 2020] [Kokoç et Altun, 2021], l'apprentissage automatique [Toti *et al*., 2020] [Hussain *et al*., 2019], [Hew et al., 2021], etc.

Notre étude de l'état de l'art relève quatre critères mesurables employés dans l'analyse des expériences d'apprentissage ; notamment, l'engagement, la satisfaction, l'évaluation de la performance des apprenants aux tests et examens et l'achèvement de la formation en ligne.

Les travaux discutés ont contribué à l'amélioration du processus d'apprentissage/ d'enseignement en ligne grâce à l'étude des expériences d'apprentissage qui a servi à renforcer les capacités des apprenants, augmenter leurs performances, résoudre quelques problèmes liés à l'EIAH, à la formation et au contenu éducatif, etc. Ces travaux montrent cependant à notre avis, quelques limites.

D'abord, de nombreuses études se sont fondées sur les données collectées à partir des réponses des apprenants à des questionnaires, enquêtes ou entretiens. Dans ce cas, la bonne qualité des données n'est pas systématiquement garantie car supposant *a priori* les réponses de l'apprenant sérieuses, objectives, fiables et dans la mesure où il consent bien à collaborer. Les données peuvent donc, facilement, être faussées ou bruitées dès que les apprenants répondent de manière non réfléchie ou commettent des erreurs de diagnostic [Melesko et Kurilovas, 2018]. La valeur scientifique de l'étude se trouve ainsi affectée.

Une seconde limite réside dans le recours à un nombre réduit de critères dans l'analyse des expériences d'apprentissage. Comme le montre le Tableau *1.2* ci-dessous, les approches proposées sont mono critère ou bi-critère tout au plus alors qu'une analyse multicritère est d'évidence pourvoyeuse d'un plus grand nombre d'information permettant une vision, une appréhension, une compréhension plus étendue et instructive des expériences d'apprentissage.

Enfin, une troisième limite que l'étude de l'état de l'art nous dévoile est l'indigence dans le traitement du sujet de l'évaluation des contenus éducatifs en ligne. Or, l'évaluation nous parait être une étape indispensable dans le processus de production du contenu éducatif de qualité. Julia et Marco notent l'importance d'une conception de qualité des contenus éducatifs en ligne, afin d'améliorer l'apprentissage et soulignent le manque d'une méthode, d'un guide, décrivant la façon de mettre en œuvre ce contenu éducatif [Julia et Marco, 2021].

Dans le cadre de cette thèse, nous proposons un système intelligent d'aide à la décision pédagogique pour assister les concepteurs dans leur tâche d'évaluation et ainsi d'amélioration des contenus éducatifs en ligne. Ce système s'appuie sur notre approche centrée apprenant dénommée MALEA (*Multicriteria Approach for Learning Experience Analysis*) ou l'approche d'analyse multicritère d'expérience d'apprentissage. L'apport de cette approche, en comparaison avec les méthodes disponibles dans la littérature, est qu'elle vise une analyse objective des expériences d'apprentissage des apprenants. A cet égard, MALEA ne se limite pas aux opinions déclarés des apprenants, mais elle utilise en plus les traces numériques d'interactions enregistrées automatiquement dans l'EIAH décrivant fidèlement les activités/le comportement des apprenants. En outre, MALEA aspire une évaluation fiable des contenus éducatifs grâce à une analyse approfondie des expériences d'apprentissage, prenant en considération, les quatre critères d'analyse proposés dans le Tableau *1.2* : engagement de l'apprenant, satisfaction de l'apprenant, évaluation de la performance de l'apprenant aux tests et examens et achèvement de la formation en ligne. L'évaluation fiable est due aussi au recours à l'apprentissage automatique permettant de se prémunir de l'imprécision du processus décisionnel humain.

Tableau 1.2 Critères utilisés dans la littérature pour analyser l'expérience d'apprentissage.

Référence	**Engagement**	**Satisfaction**	**Evaluation de la performance**	**Achèvement de la formation**
[Moubayed *et al.*, 2020]	+			
[Awidi *et al.*, 2019]	+	+		
[Asoodar *et al.*, 2016]		+		
[Krithika et Lakshmi Priya, 2016]		+		
[Hew *et al.*, 2021]		+		
[Li *et al.*, 2017]	+		+	
[Montgomery *et al.*, 2019]	+			+
[Azcona *et al.*, 2019]			+	
[Hussain *et al.*, 2019]			+	

[Klašnja-Milićević *et al.*, 2018]		+		
[Belarbi *et al.*, 2018]	+			
[Gonçalves *et al.*, 2018]			+	
[Costa *et al.*, 2019]			+	
[Bourkoukou et El Bachari, 2018]	+			
[Kaabi *et al.*, 2020]				+
[Chui *et al.*, 2020]			+	+
[Timbal, 2019]				+
[Alharbi *et al.*, 2016]			+	+
[Mubarak *et al.*, 2021]	+		+	
Notre approche visée pour l'analyse multicritère des expériences d'apprentissage (MALEA)	+	+	+	+

1.4 Conclusion

Au cours de ce chapitre, nous avons passé en revue le domaine du e-learning. Nous en avons d'abord défini le concept, les environnements informatiques pour l'apprentissage humain (EIAH) et les MOOCs. Nous avons abordé, ensuite, le phénomène d'abandon et ses facteurs. Nous nous sommes intéressés, en outre, au domaine de l'analyse des données éducatives en ligne pour son potentiel d'aide à la prise de décision pédagogique. Nous avons présenté les principales méthodes, disponibles dans la littérature, qui ont eu recours à l'analyse de l'apprentissage (*Learning Analytics*) et la fouille des données éducatives (*Educational Data Mining*) pour supporter le e-learning et améliorer les expériences d'apprentissage. Enfin, nous avons identifié et discuté les critères et les mesures utilisés dans la littérature pour analyser les expériences d'apprentissage. Ils ont été aussi recensés dans un tableau de synthèse. Celui-ci nous a permis de mettre en avant notre approche MALEA (*Multicriteria Approach for Learning Experience Analysis*) qui vise à permettre une analyse approfondie des expériences d'apprentissage considérant les quatre critères (1) engagement de l'apprenant, (2) satisfaction de l'apprenant, (2) évaluation de la performance de l'apprenant aux tests et examens et (4) achèvement de la formation en ligne. Nous adoptons

MALEA dans la quête d'une évaluation objective automatisée et fiable des contenus éducatifs en ligne.

Au chapitre suivant, nous présentons un état de l'art sur les méthodes et outils de l'apprentissage automatique supervisé et non supervisé constituant la base de nos contributions.

Chapitre 2 : L'aide à la décision dans le domaine du e-learning, vers l'utilisation de l'apprentissage automatique

2.1 Introduction

Le monde de la recherche prenant conscience de la grande quantité des données numériques aujourd'hui disponibles s'est préoccupé d'élaborer, développer, exploiter des technologies et outils permettant le traitement et l'analyse de ces données au bénéfice de différents domaines. L'apprentissage automatique (*Machine Learning*), technologie en vogue en raison de sa capacité et ses performances avérées, suscite notre intérêt. Dans ce chapitre, nous présenterons les fondements théoriques de l'apprentissage automatique, ses outils et méthodes qui sont concernés par notre méthodologie : le regroupement automatique non supervisé et la classification automatique supervisée. Le but de ce chapitre est de détailler le principe général de chaque concept et de présenter les diverses méthodes existantes que nous pourrions utiliser. Grâce à cette étude, nous allons pouvoir choisir les algorithmes qui nous semblent les plus appropriés à la mise en œuvre de notre méthodologie développée dans le chapitre 3.

2.2 Apprentissage automatique

L'apprentissage automatique appelé aussi apprentissage machine ou *Machine Learning* en anglais peut être défini comme étant l'extraction automatique des connaissances à partir des exemples de données [Harley et Sparkman, 2019], permettant aux décideurs d'obtenir plus d'informations à partir de leurs données. Il les aide à mieux comprendre ce qui s'est passé, pourquoi cela s'est passé, ce qui se passera à l'avenir, et comment y parvenir [Rajkomar et Kohane, 2019]. L'apprentissage automatique est un sous-domaine de l'intelligence artificielle [Goodfellow *et al.*, 2017] qui consiste à faire en sorte que la machine apprend et exécute des tâches sans être programmée explicitement à cet effet [Raschka et Nolet, 2020].

Dans le cadre d'une programmation qualifiable de traditionnelle, l'humain analyse les données, écrit des algorithmes dictant les instructions/règles à la machine qui les applique dans la résolution

des problèmes. Ce principe est illustré dans la Figure 2.1 ; il est ainsi question de règles exprimées dans un langage de programmation, qui agissent sur des données pour fournir des réponses [Senders *et al.*, 2018]. Plus de règles sont dictées à la machine, plus la tâche est complexe et la maintenance difficile voire insoutenable, d'où besoin de l'apprentissage automatique.

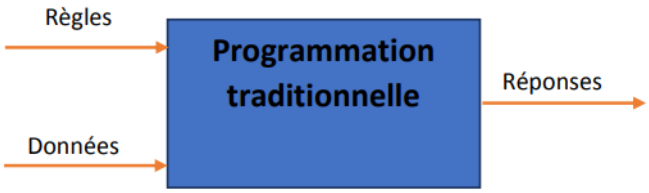

Figure 2.1 Programmation traditionnelle [Chollet, 2017] [Geisslinger, 2019]

Différemment, en apprentissage automatique, l'humain donne à la machine la capacité d'apprendre à travers les expériences dites aussi les exemples. Comme le montre la Figure 2.2, les réponses sont fournies avec les données, la machine déterminant elle-même les règles [Senders *et al.*, 2018].

Figure 2.2 Apprentissage automatique [Chollet, 2017] [Geisslinger, 2019]

L'apprentissage automatique est capable de traiter des problèmes complexes où la programmation qualifiable de traditionnelle montre ses limites. Si nous voulons, par exemple, construire un programme qui conduit une voiture, nous allons nous trouver devant un nombre infini de cas possibles à traiter, ce qui rend la tâche très difficile voire impossible. Le *Machine Learning* traite cette problématique autrement. On considère qu'il n'y a plus besoin de décrire quoi faire, parce que le programme va apprendre par lui-même comment conduire en observant des expérimentations [Stilgoe, 2018]. Un algorithme d'apprentissage peut être plus performant que ses programmeurs humains. Des systèmes basés sur des algorithmes d'apprentissage automatique sont devenus des champions du monde dans des jeux comme les dames et les échecs [Strogatz, 2018] [Risi et Preuss, 2020]. Cela serait impossible si les programmes ne faisaient que ce qu'on leur avait explicitement dicté leurs actions.

Dans cette section nous exposons, d'abord, les critères de choix de l'apprentissage automatique pour résoudre notre problématique. Ensuite, nous définissons les notions fondamentales de l'apprentissage automatique ainsi que ses types et son processus général.

2.2.1 Notions fondamentales de l'apprentissage automatique

En apprentissage automatique, la machine apprend à partir des exemples de données comment exécuter des tâches [Garcia *et al*., 2018]. Pour comprendre comment la machine apprend, nous expliquons les trois notions fondamentales de l'apprentissage automatique [Simeone, 2018] [Qiu *et al*., 2016] [Mahesh, 2020] : données, caractéristiques et algorithme.

- **Données**

Pour apprendre, la machine a besoin d'exemples. Les exemples sont regroupés dans un tableau de données appelé ensemble de données (*dataset*). Chaque ligne du tableau de données est appelée observation. Manuellement ou automatiquement, il est extrêmement difficile de rassembler une bonne collection de données. Les données sont si importantes que les entreprises peuvent même révéler leurs algorithmes, mais rarement leurs ensembles de données. Généralement, la collecte des données est semi-automatisée [Paullada *et al*., 2021].

- **Caractéristiques (Features)**

Les caractéristiques sont également connues sous le nom de paramètres ou de variables. Il peut s'agir du genre de l'apprenant, de son âge, son pays, son niveau d'engagement dans le cours en ligne, etc. En d'autres termes, ce sont les facteurs qu'une machine doit examiner. Lorsque les données sont stockées dans des tableaux, les caractéristiques sont les noms de colonnes. Dans le cas d'un grand nombre de caractéristiques, il faut choisir uniquement les caractéristiques les plus importantes servant à la résolution du problème donné [Cai *et al*., 2018]. Par exemple si nous désirons apprendre à une machine à distinguer une voiture d'une bicyclette, la caractéristique nombre de roues est très importante contrairement à la caractéristique couleur.

- **Algorithme**

L'apprentissage automatique se fait à travers des algorithmes qui sont catégorisés selon le type d'apprentissage [Mahesh, 2020]. Ainsi, tout problème peut être résolu différemment. Il peut survenir que la qualité et/ou la quantité des données ne permettent pas d'arriver aux meilleurs résultats, c'est le choix de l'algorithme qui déterminera la performance du modèle d'apprentissage

automatique [Sharma et Kumar, 2017], raison pour laquelle il est recommandé de tester plusieurs algorithmes, comparer les résultats et adopter le plus pertinent. Nous exposons dans les sections 2.3.3 et 2.4.3 des exemples d'algorithmes les plus réputés et employés pour l'apprentissage supervisé et non supervisé.

2.2.2 Types d'apprentissage automatique

Dans cette section, nous exposons les trois types principaux d'apprentissage automatique, en l'occurrence l'apprentissage automatique supervisé, l'apprentissage automatique non supervisé et l'apprentissage automatique par renforcement [Portugal *et al*., 2018], bien qu'il existe un quatrième type qui s'appelle l'apprentissage semi-supervisé [Sarker, 2021] et qui utilise ce qu'il y a de mieux dans les deux premiers types pour créer une sorte d'hybride ; tel est le cas de cette thèse.

- **Apprentissage automatique supervisé**

Les humains apprennent à partir des exemples. Si des parents souhaitent apprendre à leur enfant à reconnaitre une voiture, ils doivent lui montrer plusieurs exemples de voitures. L'apprentissage supervisé s'inspire de ce mode de fonctionnement d'apprentissage humain. Dans l'apprentissage supervisé c'est l'expert en science des données (*Data Scientist*) qui joue le rôle des parents. Celui-ci fournit à la machine des exemples de données labélisées ou étiquetées qu'elle doit étudier [Garcia *et al*., 2018]. Ces exemples sont regroupés dans un tableau de données qui contient deux types de variables :

- une ou plusieurs variables d'entrée *x* appelées variables indépendantes. Ce sont les caractéristiques ou les facteurs (*features*) qui viennent influencer la valeur de la variable dépendante de sortie *y*.
- une variable de sortie *y* dite variable dépendante, étiquette ou label (*target*). Elle constitue l'objectif attendu. C'est la réponse que nous voulons que la machine apprenne à prédire.

Grâce à ce tableau de données la machine sera capable d'établir des règles entre les variables d'entrée et celles de sortie, afin d'associer une nouvelle variable d'entrée non-étiquetée à une variable de sortie [Schrider et Kern, 2018]. Ces règles sont appelés modèle. Ainsi, l'apprentissage supervisé consiste à laisser une machine développer un modèle *f* à partir d'un ensemble de données labélisées (*x*, *y*) tel que $y=f(x)$. Il appartient à l'expert en science des données de choisir le type du modèle. Quant à la machine, elle doit trouver/apprendre les valeurs des paramètres du modèle qui mènent aux meilleurs résultats. Selon la définition de Tom Mitchell [Mitchell, 1997], une machine

apprend si sa performance à réaliser une tâche donnée s'améliore avec l'expérience, c'est-à-dire avec le nombre d'exemples considérés.

- **Apprentissage automatique non supervisé**

Contrairement à l'apprentissage supervisé qui tente de trouver un modèle depuis des données labellisées, l'apprentissage non supervisé se fait d'une façon totalement autonome. La machine reçoit des données qui ne sont pas labélisées, c'est-à-dire sans les réponses attendues. Un algorithme d'apprentissage non supervisé cherche par lui-même une structuration ou des patterns dans les données fournies [Oliver *et al.*, 2018].

- **Apprentissage automatique par renforcement**

Dans les problèmes d'apprentissage par renforcement, un agent interagit avec l'environnement dans le but de trouver la solution optimale. Ce type d'apprentissage se distingue des problèmes supervisés et non supervisés par son côté interactif et itératif [Kaelbling et Moore, 1996]: l'agent explore en essayant plusieurs solutions, il observe la réaction de l'environnement puis adapte son comportement afin de trouver la meilleure stratégie et exploite le résultat de ses explorations [François-Lavet *et al.*, 2018]. L'équilibre entre les phases d'exploration et d'exploitation est l'un des concepts clés de ce type de problèmes. Le scenario typique d'apprentissage par renforcement est le suivant : l'agent effectue une action sur son environnement, l'action est interprétée en une récompense et une représentation de l'état nouveau est transmise à l'agent tel que illustré dans la Figure 2.3 [Hew *et al.*, 2020].

Grâce à ce type d'apprentissage, un robot peut apprendre à marcher dans un milieu complexe et incertain tel est le cas de la voiture autonome, ou à accomplir une tâche bien spécifique comme la planification, ou le jeu d'échecs, etc.

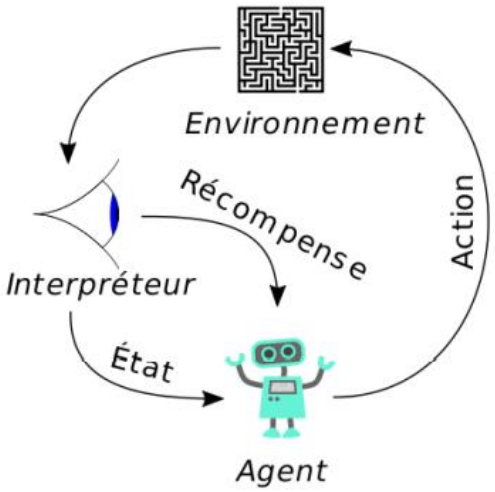

Figure 2.3 Scenario typique de l'apprentissage par renforcement [Reinforcement Learning, 2017]

2.2.3 Processus général d'apprentissage automatique

Le processus général d'apprentissage automatique se compose principalement de sept étapes [Zhu *et al.*, 2019], comme illustré dans la Figure 2.4 ci-dessous.

La première étape est la collecte des données. Il s'agit d'une tâche importante car elle déterminera la qualité du modèle. Pratiquement, les données recueillies sont souvent non structurées, contiennent des bruits ou doivent prendre d'autres formes pour pouvoir être utilisables par l'algorithme d'apprentissage automatique. Par conséquent, une étape de nettoyage et de prétraitement des données s'impose. Ensuite, la construction du modèle d'apprentissage automatique peut commencer. Celle-ci se déroule en deux étapes : (1) l'ingénierie des caractéristiques dans laquelle s'opère la sélection des plus pertinentes des caractéristiques, (2) puis le choix de l'algorithme d'apprentissage automatique répondant au problème en question. Ces deux choix décideront de la performance du modèle. La cinquième étape est consacrée à l'apprentissage. Au cours de la sixième étape, la performance du modèle doit être testée. Parfois, il est possible de revenir en arrière et améliorer la phase d'apprentissage. La dernière étape est le résultat donné par l'apprentissage. Il peut s'agir d'une prédiction ou d'une inférence.

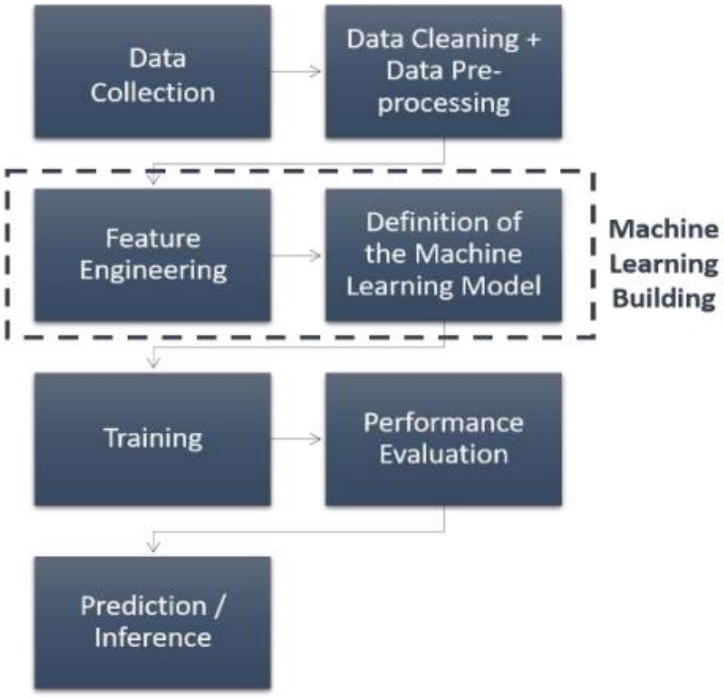

Figure 2.4 Processus général d'apprentissage automatique [Mourali *et al*., 2020b]

2.3 Apprentissage automatique supervisé

Dans cette section seront abordés les types de problèmes, les outils d'évaluation de performance et les algorithmes les plus notoires de l'apprentissage automatique supervisé.

2.3.1 Problèmes résolus par l'apprentissage automatique supervisé

L'apprentissage automatique supervisé peut résoudre deux types de problèmes : la régression et la classification [Nasteski, 2017].

- **Classification**

La classification est peut-être le problème d'apprentissage automatique supervisé le plus courant. Elle consiste à faire entrainer un modèle de manière à associer une valeur de sortie discrète *y* à une ou plusieurs valeurs d'entrée *x*. Cette valeur de sortie correspond à une classe ou une étiquette, reflétant le terme de classification. Il existe diverses applications qui utilisent la classification, citons le diagnostic médical tel que l'identification des cellules cancéreuses [Yue *et al*., 2018] [Dhahri *et al*., 2019], le marketing ciblé [Hair et Sarstedt, 2021] [Salminen *et al*., 2019], la

détection de spam [Makkar *et al*., 2020] [GuangJun *et al*., 2020], la prédiction du risque de crédit [Bussmann *et al*., 2021] [Uthayakumar *et al*., 2020] [Bhatore *et al*., 2020], la classification des documents [Kadhim, 2019] [Kim *et al*., 2019] et l'analyse des sentiments [Mitra, 2020] [Rathi *et al*., 2018], pour n'en citer que quelques-uns.

- **Régression**

La régression est une tâche d'apprentissage automatique supervisée qui permet de prédire des valeurs numériques continues en fonction d'une ou de plusieurs valeurs d'entrée. Parmi les cas d'utilisation les plus courants, citons la prévision des ventes et de la demande [Kohli *et al*., 2021] [Catal *et al*., 2019], la prévision du cours des actions, de l'immobilier ou des marchandises [Henrique *et al*., 2018] [Phaladisailoed et Numnonda, 2018], et la prévision météorologique [Maulud et Abdulazeez, 2020] [Hossain et Mahmood, 2020], pour n'en citer que quelques-uns.

Dans le cadre de cette thèse nous nous proposons d'aider les concepteurs pédagogiques à évaluer leurs contenus éducatifs en ligne afin de pouvoir les améliorer. Nous opterons pour une classification binaire des contenus éducatifs en ligne dans le but d'identifier les contenus qui nécessitent des améliorations. Ainsi, nous souhaitons de faire apprendre à la machine à prédire la classe d'un contenu éducatif donné. La classe prédite, variable de sortie, peut avoir deux catégories : *réussi* ou *à améliorer*.

2.3.2 Evaluation de performance d'un modèle de classification

Il est nécessaire de savoir évaluer les performances d'un modèle d'apprentissage automatique afin de pouvoir choisir le bon modèle pour un problème déterminé. L'une des manières les plus répandues pour mesurer la performance d'un modèle de classification est la matrice de confusion [Boutaba *et al*., 2018], également connue sous le nom de tableau de contingence. Pour calculer une matrice de confusion, il faut disposer d'un ensemble de données de test labélisé. La matrice de confusion résume les résultats de prédiction pour un problème de classification particulier. Elle montre à quel point un certain modèle peut être confus en révélant le nombre de prédictions justes et fausses [Haghighi *et al*., 2018]. En comparant l'étiquette connue et la classe prédite pour chaque point de données, les résultats peuvent être classés dans l'une des quatre catégories suivantes :

- Vrais Positif (*True Positive*) lorsque la classe réelle et la classe estimée sont toutes les deux positives ; par exemple un contenu éducatif réussi est estimé réussi.

- Vrais Négatif (*True Negative*) lorsque la classe réelle et la classe estimée sont toutes les deux négatives ; par exemple un contenu éducatif qui nécessite des améliorations est estimé un contenu à améliorer.
- Faux Positif (*False Positive*) lorsque la classe réelle est négative mais que la classe estimée est positive ce qui représente une erreur : par exemple un contenu éducatif qui nécessite des améliorations est estimé réussi.
- Faux Négatif (*False Negative*) lorsque la classe réelle est positive mais que la classe estimée est négative ce qui représente une erreur ; par exemple un contenu éducatif réussi est estimé nécessitant des améliorations.

Dans le cas d'une classification binaire, la matrice de confusion sera une matrice de 2 par 2, avec quatre valeurs, tel que décrit dans le Tableau 2.1 suivant [Boutaba *et al.*, 2018] [Luque *et al.*, 2019].

Tableau 2.1 Matrice de confusion pour une classification binaire [Boutaba *et al.*, 2018] [Luque *et al.*, 2019]

	Classe prédite négative	**Classe prédite positive**
Classe réelle négative	Vrai Négatif (VN)	Faux Positif (FP)
Classe réelle positive	Faux Négatif (FN)	Vrai Positif (VP)

L'avantage de la matrice de confusion est qu'elle est très simple à lire et à comprendre [Hasnain *et al.*, 2020]. En outre, elle permet non seulement de savoir quelles sont les erreurs de prédiction commises, mais surtout le type de ces erreurs. La matrice de confusion peut, en effet, être utilisée pour des mesures plus approfondies comme l'exactitude, la précision, le rappel, la spécificité et le score F1 [Hussain *et al.*, 2019]. Ces mesures, qui sont toutes décrites ci-après, permettent d'obtenir une évaluation de la qualité du modèle de classification et d'identifier les tendances qui peuvent aider à l'améliorer. En général, pour chaque application, l'expert en science des données doit choisir soigneusement une mesure à utiliser qui dépend de son problème. Le choix de la mesure peut influencer la manière dont la performance est évaluée et interprétée.

- **Exactitude (Accuracy)**

L'exactitude permet de connaitre la proportion des prédictions correctes fournies par le modèle par rapport à toutes les prédictions [Sezer et Altan, 2021]. Des valeurs élevées de cette mesure sont souhaitables. L'exactitude représente le ratio entre le nombre de prédictions correctes et le nombre total de prédictions [Lopez-Bernal *et al*., 2021]. Ceci peut être calculé en utilisant les valeurs de la matrice de confusion et en appliquant la formule suivante :

$$Exactitude = \frac{(\mathrm{VP} + \mathrm{VN})}{(\mathrm{VP} + \mathrm{VN} + \mathrm{FP} + \mathrm{FN})}$$

L'exactitude n'est pas la mesure idéale dans les situations où l'ensemble de données est déséquilibré [Juba et Le, 2019]. Pour l'illustrer à l'aide d'un exemple, considérons une tâche de classification avec 90 observations négatives et 10 observations positives. Classer toutes les observations comme négatives donne un score d'exactitude élevé de 0,90. Cette mesure ne permet pas de déterminer le nombre des vrais positifs que le modèle peut identifier. La précision et le rappel sont de meilleures mesures pour évaluer les modèles entrainés avec des données déséquilibrées.

- **Précision (Precision)**

La précision est la capacité d'un modèle à identifier uniquement les objets pertinents. Il s'agit du pourcentage de prédictions positives correctes [Padilla *et al*., 2020]. La précision correspond au ratio entre le nombre de prédictions (classifications) positives correctes (VP) et le nombre total de prédictions positives [Sezer et Altan, 2021]. Pour notre problème de classification des contenus éducatifs ce serait la mesure des contenus éducatifs en ligne que nous identifions correctement comme réussi parmi tous les contenus éducatifs réellement réussis. Elle est calculée à l'aide de la formule suivante :

$$Précision = \frac{\mathrm{VP}}{(\mathrm{VP} + \mathrm{FP})}$$

Cette mesure donne une indication sur le taux des faux positifs, ceux classés comme positifs par erreur. Dans notre cas se sont les contenus éducatifs prédits réussis mais qui nécessitent, en réalité, des améliorations. Il est important de minimiser les faux positifs pour améliorer la qualité des contenus éducatifs diffusés sur les EIAH. La précision est donc une bonne mesure à utiliser dans les situations où le nombre des faux positifs est élevé.

- **Rappel (Recall/ sensitivity)**

Le rappel ou sensibilité désigne la proportion des valeurs positives prédites avec précision [Sezer et Altan, 2021]. Cette mesure correspond donc au ratio entre le nombre de prédictions positives correctes et le nombre total de classifications de classe positive [Padilla *et al*., 2020]. Ainsi dans notre cas, pour tous les contenus éducatifs réussis, le rappel signale combien notre modèle a correctement identifié de contenus éducatifs réussis. On peut utiliser la formule suivante :

$$Rappel = \frac{\text{VP}}{(\text{FN} + \text{VP})}$$

Cette mesure donne une indication sur la part des faux négatifs ; dans notre cas les contenus éducatifs réussis qui n'ont pas été détectés par le modèle de classification. Quand un contenu éducatif est réussi mais a été prédit par le modèle de classification ayant besoin d'amélioration, le travail de conception pédagogique risque de se prolonger indûment et par conséquent il devient improductif et non rentable.

- **Spécificité (Specificity)**

La spécificité correspond au nombre de classes négatives prédites par le modèle [Sezer et Altan, 2021]. Cette mesure est déterminée par le ratio entre le nombre de prédictions négatives correctes et le nombre total de prédictions négatives. Elle peut se calculer de la manière suivante :

$$Spécificité = \frac{\text{VN}}{(\text{VN} + \text{FP})}$$

Cette mesure est très importante dans notre cas. Elle indique à quel point notre modèle est capable de détecter les contenus éducatifs qui nécessitent des améliorations. Augmenter la spécificité permet d'optimiser la détection des contenus éducatifs qui nécessitent des améliorations.

- **Score F1 (F1 score)**

Le score F1 est la moyenne harmonique de la précision et du rappel [Gaussier et Yvon, 2011] [Clavel, 2019]. Cette mesure est utilisée lorsque nous nous trouvons face au dilemme de devoir faire le choix entre l'élévation de la valeur du rappel ou de celle de la précision [Lipton, 2014]. Le recours à la moyenne harmonique au lieu de la moyenne arithmétique constitue une prévention contre les valeurs extrêmes de rappel et de précision. Dans le cas où la précision est égale à 1 et le rappel est égal à 0, le modèle donnera toujours et systématiquement la même prédiction (l'une ou

l'autre des deux classes). La moyenne arithmétique obtenue serait égale à 0,5, valeur exagérée pour un modèle inepte qui ignore l'entrée et prédit simplement l'une des classes en sortie. En revanche la moyenne arithmétique serait égale à 0, évaluation de performance plus proche de la réalité. La moyenne harmonique est donnée par la formule suivante, permettant la possibilité d'accorder un poids plus élevé à la précision ou au rappel grâce au paramètre ajustable β [Sasaki, 2007]:

$$Score\ F_\beta = (1+\beta^2) * \frac{\text{Rappel} * \text{Précision}}{(\beta^2 * \text{Rappel} + \text{Précision})}$$

Le score F1 est souvent utilisé pour évaluer la performance d'un modèle de classification. Il est particulièrement utile lorsque l'identification des vrais négatifs est relativement peu importante, car le taux de vrais négatifs n'est pas compris dans le calcul de la précision ou du rappel [Takahashi *et al.*, 2022].

$$Score\ F1 = 2 * \frac{\text{Rappel} * \text{Précision}}{(\text{Rappel} + \text{Précision})}$$

- **Validation croisée (Cross Validation)**

La validation croisée est souvent utilisée pour la validation d'un modèle de classification en raison de la faible quantité de données disponibles. Le principe de la validation croisée est illustré dans la Figure 2.5. Il consiste à partitionner l'ensemble des données en *k* blocs de même taille. Parmi les *k* blocs, un bloc est conservé comme base de test pour valider le modèle, et les autres *k*-1 blocs sont utilisés comme base d'apprentissage. Ce processus est ensuite répété *k* fois (itérations). La validation croisée a l'avantage de permettre une utilisation équilibrée des données à la fois pour l'apprentissage (base d'apprentissage) et la validation (base de test) [Refaeilzadeh and Liu, 2009]. La performance d'un modèle est donnée par le calcul de la moyenne des performances.

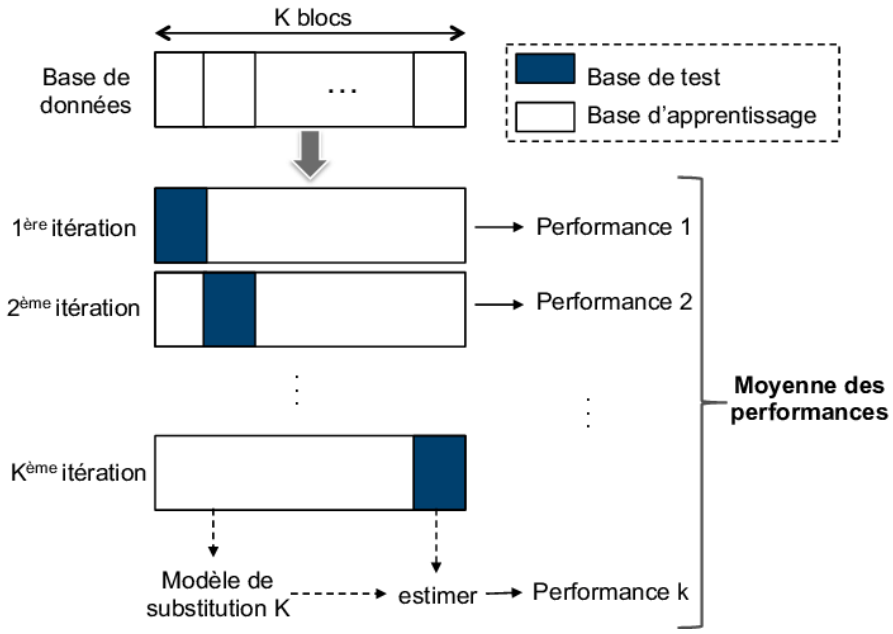

Figure 2.5 Principe de la validation croisée [Laqrichi, 2015]

2.3.3 Algorithmes incontournables de classification

Il existe de nombreux algorithmes de classification largement appliqués. Nous faisons le choix de présenter le perceptron, l'adaline avec la descente de gradient et la descente de gradient stochastique, la machine à vecteur de support et la régression logistique en raison de leur utilisation fréquente spécifiquement dans la classification binaire et de bons résultats obtenus dans plusieurs domaines.

- **Perceptron**

Le perceptron est un algorithme de classification binaire [Sagheer *et al.*, 2019]. Il peut être considéré comme l'un des premiers et l'un des plus simples types de réseaux neuronaux artificiels. Dans sa version simplifiée, le perceptron possède deux couches : la couche d'entrée *X* et la couche de sortie *Y*. La couche d'entrée correspond à l'ensemble des données d'apprentissage. La couche de sortie correspond à la prédiction de la classe de chaque donnée de l'ensemble des données d'apprentissage.

Soit $S= \{(x_i, y_i)\}_{i=1}^{n}$ l'ensemble de données d'apprentissage contenant *n* observations,

Avec

$x_i \in \mathbb{R}^d$, est la *i*-ème observation représentée par le vecteur de caractéristiques de dimension *d*

$y_i \in \{-1, 1\}$, est le label/classe de la *i*-ème observation, les valeurs -1 et 1 présentent respectivement la classe négative et la classe positive.

Comme montré dans la Figure 2.6, un perceptron est décrit par un vecteur de poids $w(w_1, \dots, w_d)$, un biais w_0 et un seuil d'activation Θ. Un perceptron doit apprendre le vecteur *w* en satisfaisant les conditions suivantes :

$$(x_1, \dots, x_d) \mapsto y = \begin{cases} +1 \; si \; w_0 + x_1 w_1 + x_2 w_2 + \cdots + x_d w_d \geq \Theta \\ -1 \; si \; w_0 + x_1 w_1 + x_2 w_2 + \cdots + x_d w_d \leq \Theta \end{cases}$$

L'idée de l'algorithme du perceptron est d'initialiser *w* au vecteur nul, itérer un nombre de fois fixé *a priori* ou jusqu'à convergence, sur les données d'apprentissage, et ajuster le vecteur de pondération w à chaque fois qu'une donnée est mal classée [Lopez-Bernal *et al.*, 2021].

La couche de sortie reçoit la somme pondérée par des poids. Le réseau est déclenché par la réception d'une information en entrée. Le traitement de la donnée dans ce réseau se fait entre la couche d'entrée et la couche de sortie qui sont toutes reliées entre elles [Lopez-Bernal *et al.*, 2021].

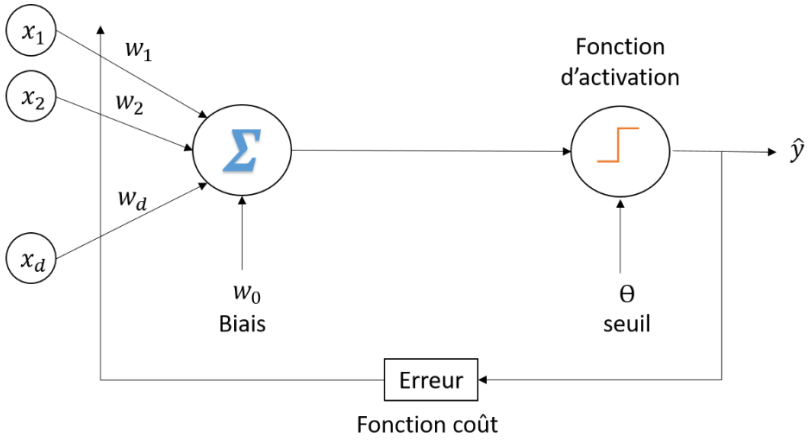

Figure 2.6 Illustration de l'algorithme perceptron

A l'instar de la régression logistique (décrite en 2.3 et 3.4), le perceptron peut apprendre rapidement grâce à son recours à l'algorithme d'optimisation de descente de gradient stochastique à situer la séparation linéaire permettant la classification binaire. Cependant, contrairement à la régression logistique, le perceptron n'indique pas la probabilité d'appartenance d'un élément à telle ou telle classe.

- **Adaline (Adaptive Linear Neuron)**

Adaline est un réseau de neurones simple qui permet de faire des classifications binaires [Deepa *et al.*, 2021]. Comme le perceptron, adaline est caractérisé par un vecteur de poids, un biais, une fonction de sommation et une fonction d'activation. En termes d'architecture, il existe une différence notable entre le perceptron et l'adaline. Le perceptron utilise une fonction d'activation basée sur le seuil alors que l'adaline utilise une fonction d'activation linéaire, ce qui a des implications sur la manière avec laquelle l'algorithme apprend son vecteur de poids. Le perceptron met à jour ses poids en changeant leurs valeurs lorsqu'un ou plusieurs échantillons de données sont prédits de manière incorrecte pendant l'apprentissage. Pour l'adaline, les mises à jour des poids se font sur la base de l'algorithme de descente de gradient, ou l'algorithme de descente de gradient stochastique [Deepa *et al.*, 2021], tel que montré en Figure 2.7. Ces mises à jours peuvent se produire même si tous les exemples sont correctement prédits tout au long de la phase de l'apprentissage. Cela est dû au fait que la variable de sortie $\hat{Y}$ n'est plus une variable catégorielle mais une variable continue représentant la probabilité d'appartenance à telle ou telle classe et que la minimisation de l'erreur est effectuée avec une fonction de coût pour déterminer les meilleurs poids. C'est l'atout majeur de l'algorithme adaline.

Comme la sortie de cette fonction est une variable continue et la sortie attendue pour un problème de classification est une variable catégorielle, adaline utilise un quantificateur pour transformer la sortie en l'une des deux valeurs (-1 ou 1). Cette fonction d'activation supplémentaire est appelée fonction de signe.

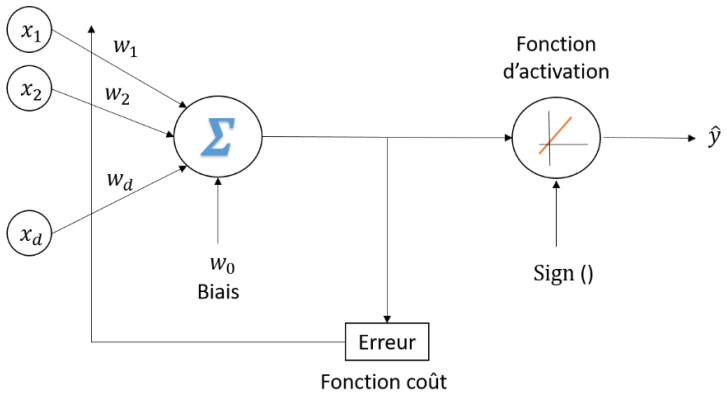

Figure 2.7 Illustration de l'algorithme adaline

Puisque adaline applique souvent l'algorithme de descente de gradient ou l'algorithme de descente de gradient stochastique, il convient de présenter ces deux derniers. La fonction coût de adaline est la somme des distances entre la valeur attendue et la valeur prédite par l'algorithme de classification sur l'ensemble des données disponibles. Cette fonction de coût est donnée par :

$$C(w) = \frac{1}{n}\sum_{i=1}^{n} C_i(\mathrm{w})$$

Selon la définition de Tom Mitchell, une machine apprend si ses performances à réaliser une tâche donnée s'améliorent avec l'expérience, c'est-à-dire avec le nombre d'exemples considérés [Mitchell, 1997]. L'algorithme de descente de gradient est un algorithme d'optimisation utilisé pour trouver l'ensemble de poids *w* qui minimise l'erreur globale des prédictions du réseau [Alpaydin, 2020]. Pour y parvenir, soit *C(w)* la fonction coût, l'algorithme de descente de gradient essaie de manière itérative de trouver le minimum global de cette fonction en calculant la dérivée partielle par rapport aux poids. La valeur des poids w qui minimise la fonction coût *C(w)* est trouvée si :

$$\frac{\partial C(w)}{\partial w} = 0$$

Pour satisfaire cette condition l'algorithme de descente de gradient se base sur la méthode de la descente locale qui consiste à modifier itérativement w pour diminuer $C(w)$ jusqu'à atteindre un minimum local. Ainsi la formule de descente de gradient est la suivante [Kuncheva, 2014] :

$$w^{k+1} = w^k - p_k \frac{\partial C(w^k)}{\partial w^k}$$

avec w^k représentant le vecteur des poids à l'itération k, et p_k un scalaire appelé pas de gradient *(learning rate)*, qui peut être fixé, adaptatif, ou déterminé par un échéancier. Si le pas de gradient est trop grand alors les mises à jour font augmenter le coût au lieu de le diminuer. Si le pas de gradient est trop petit, la convergence est plus lente. La Figure 2.8 décrit la manière dont la fonction coût *C(w)* varie en fonction d'une valeur de poids en utilisant l'algorithme de descente de gradient.

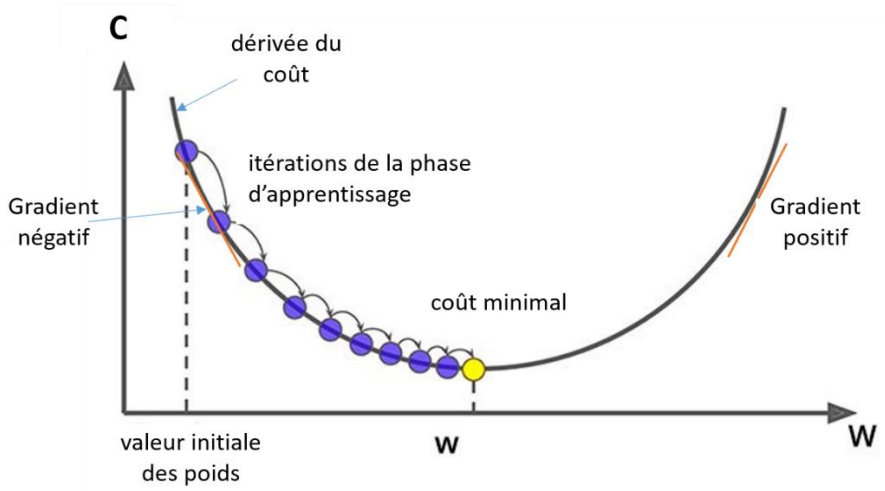

Figure 2.8 Apprentissage avec la descente de gradient

En apprentissage, le nombre *n* de données utilisées peut être très grand. Dans ce cas le calcul peut être couteux en termes de temps. L'algorithme de descente de gradient stochastique procède à une simplification. Il traite un lot de données tirées aléatoirement à chaque itération. Chaque fonction de coût minimisée de cette manière est une approximation de la fonction objective globale. Par conséquent, les poids sont mis à jour même après une itération dans laquelle un seul exemple a été traité. C'est donc plus rapide que la descente de gradient ordinaire.

- **Machines à vecteurs de support (Support Vector Machines)**

Les machines à vecteurs de support (SVM) appelées aussi séparateurs à vaste marge [Hasan et Boris, 2006] représentent une méthode populaire de classification binaire [Chauhan *et al*., 2019]. Son principe de base est illustré dans la Figure 2.9 : un algorithme SVM doit trouver la frontière/hyperplan optimal de telle façon que la distance/ marge entre les différents groupes de données et la frontière qui les sépare soit maximale. Les points de données les plus proches de la limite de classification sont appelés vecteurs de support [Mekonnen *et al*., 2019].

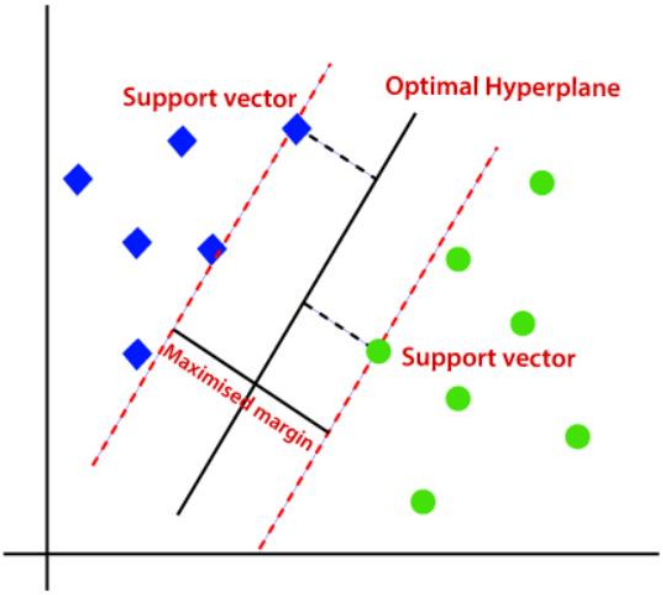

Figure 2.9 Illustration de la machine à vecteur de support [Baghaee *et al.*, 2019]

Pour trouver l'hyperplan optimal, la machine à vecteur de support se base souvent sur l'utilisation des noyaux. Un noyau est une fonction mathématique φ permettant de séparer linéairement les données en les projetant dans un espace vectoriel de plus grande dimension (*feature space*) tel que schématisé dans la Figure 2.10.

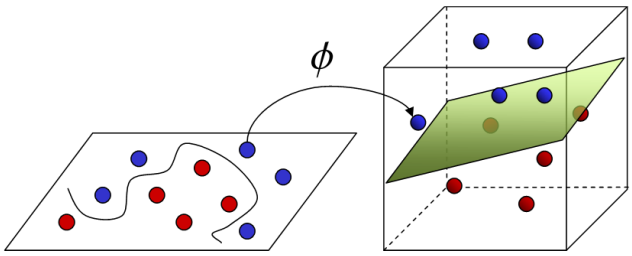

Figure 2.10 Projection des données en dimension 3 [Yue *et al.*, 2010]

Soit :

$H : (w \cdot x) + b$ l'hyperplan qui satisfait les conditions suivantes, avec xi étant la i-ème observation :

$$w \cdot xi + b \geq 1 \; si \; yi = 1$$

$$w \cdot xi + b \leq -1 \; si \; yi = -1$$

L'hyperplan optimal *HO* est l'hyperplan qui maximise la marge *M*. La marge *M* représente la plus petite distance entre les différentes données des deux classes et l'hyperplan. Maximiser la marge *M* est équivalent à maximiser la somme des distances des deux classes par rapport à l'hyperplan. Ainsi, la marge est calculée de la manière suivante [Kharroubi, 2002] :

$$M = min_{X_i|y_i=1} \frac{w \cdot x + b}{||w||} - max_{X_i|y_i=-1} \frac{w \cdot x + b}{||w||}$$

$$= \frac{1}{||w||} - \frac{-1}{||w||}$$

$$= \frac{2}{||w||}$$

Trouver l'hyperplan optimal *HO* revient donc à maximiser $\frac{2}{||w||}$ et par la suite à minimiser $||w||$.

L'un des principaux avantages de cet algorithme réside dans le nombre très réduit de ses hyper-paramètres qui se limitent au choix de la technique de régularisation (lasso par exemple) et au choix du noyau (noyau polynomial par exemple). Cet algorithme est reconnu pour sa capacité de traiter des données de grandes dimensions, par ses garanties théoriques et ses performances dans la pratique [Yue *et al.*, 2019]. Requérant un faible nombre de paramètres, les SVM sont appréciées pour leur simplicité d'usage.

Dans le secteur du e-learning, Azcona *et al*. ont construit un modèle prédictif à l'aide des SVM. Ce modèle est capable de distinguer les apprenants en besoin d'aide lors de l'apprentissage de ceux qui risquent d'échouer à leur prochaine évaluation [Azcona *et al*., 2019].

Sur la base des notes d'évaluation, Chui *et al*. ont eu, aussi, recours aux machines à vecteurs de support (SVM) pour prédire les apprenants marginaux ceux susceptibles d'échouer et ceux à risque certain d'échouer à travers leurs notes d'évaluation [Chui *et al*., 2020]. Ces deux travaux de recherche ont été détaillés dans la section 1.3.1.

- **Régression logistique (Logistic Regression)**

La régression logistique est un classificateur linéaire binaire qui prédit les probabilités [Lever et Altman, 2016]. Il utilise une fonction logistique (sigmoïde) pour transformer sa sortie en une valeur de probabilité qui peut être associée à deux classes. Un classificateur linéaire tel que la régression logistique convient lorsque les données ont une limite de décision claire.

En rapport avec l'amélioration de l'apprentissage et l'acquisition des nouvelles connaissances et compétence chez les apprenants, Hussain *et al*. visent à prédire les performances des étudiants ainsi que les difficultés qu'ils peuvent rencontrer dans une session de cours de conception numérique [Hussain *et al*., 2019]. Alharbi *et al*. visent à identifier, le plus en amont possible, les apprenants à risquent de performance académique faible [Alharbi *et al*., 2016]. Dans la section 1.3.1 nous avons fourni plus d'informations sur ces deux travaux.

Nous recourrons à cet algorithme afin de classifier les contenus éducatifs en ligne. Dans le chapitre 3 nous précisons les critères de choix de cet algorithme et détaillons son fonctionnement. Dans le chapitre 4, nous concrétisons la régression logistique sur un cas d'étude réel et présentons les résultats de son expérimentation.

2.4 Apprentissage automatique non supervisé

Dans cette section seront évoqués les types de problèmes résolubles, les outils d'évaluation de performance et les algorithmes les plus notoires de l'apprentissage automatique non supervisé.

2.4.1 Problèmes résolus par l'apprentissage automatique non supervisé

L'apprentissage automatique non supervisé peut résoudre trois types de problèmes : le regroupement, la détection d'anomalie et la réduction de dimensions. Nous exposons ci-dessous les trois types de problèmes en accordant une attention particulière au regroupement (*clustering*), objet de l'une de nos contributions.

- **Regroupement (*Clustering*)**

Le regroupement connu sous le nom de classification non supervisée ou *clustering* est le type de problèmes le plus répandu en apprentissage automatique non supervisé [Madhulatha, 2012] [Kassambara, 2017]. Il consiste à faire apprendre à la machine à classer les observations de données selon leur ressemblance. Ainsi la machine sera capable de regrouper des observations non étiquetées dans des groupes similaires. Parmi les cas d'utilisation populaires du *clustering*, nous pouvons citer la classification des documents, des photos, des tweets, la segmentation des utilisateurs, etc. Le regroupement est également utilisé pour générer des données d'apprentissage pour les classificateurs dans les cas où les données d'apprentissage sont indisponibles, comme dans notre cas d'étude [Hofmann, 2001]. Dans cette thèse nous nous intéressons à l'évaluation des contenus éducatifs en ligne. Pour ce faire nous avons besoin d'apprendre à la machine à classer les

contenus éducatifs en contenus réussis et en contenus nécessitant des améliorations. Pour entrainer la machine, et comme nous le détaillerons dans les chapitres 3 et 4, nous mettrons à sa disposition un ensemble de données/ d'observations labélisées. Une observation correspond à un exemple de contenu éducatif en ligne. Une observation est composée de l'ensemble des caractéristiques du contenu éducatif en ligne et de son label. Le label indique la classe du contenu éducatif (réussi ou nécessitant des améliorations). Nous utiliserons un regroupement pour la labélisation de l'ensemble de nos données.

Il existe plusieurs méthodes et algorithmes de regroupement comme le K-means appelé aussi k-moyenne, la propagation d'affinité (*Affinity Propagation*) [Laureano *et al*., 2020] [Wang *et al*., 2019] [Karga et Satratzemi, 2018], le partitionnement spectral (*Spectral Clustering*) [Mengoni *et al*., 2018] [Cavallari *et al*., 2017] et le regroupement aggloméré (*Agglomerative Clustering*) [Hussain *et al*., 2018] [Bharara *et al*., 2018] [Hassel et Ridout, 2018].

Les méthodes de regroupement sont divisées en trois grandes familles : le regroupement hiérarchique, le regroupement par partitionnement [Saxena *et al*., 2017] et les cartes auto organisatrices [Kohonen, 1982].

La première famille, celle d'algorithmes de regroupement hiérarchique, est divisée en deux branches : les algorithmes ascendants dits aussi agglomérés et les algorithmes descendants. Les algorithmes agglomérés commencent par considérer tout d'abord que chaque observation est un groupe. Ensuite, les deux groupes les plus proches sont agglomérés en un seul groupe. Cette étape est répétée jusqu'à ce que toutes les observations soient regroupées en un seul groupe. Contrairement aux algorithmes agglomérés, les algorithmes descendants démarrent en rassemblant toutes les observations dans un seul groupe. Puis, les observations sont divisées à chaque étape selon un critère jusqu'à obtenir autant de groupes que d'observations.

La seconde famille d'algorithmes de regroupement est celle des algorithmes de regroupement par partitionnement. Ces algorithmes rassemblent les observations en plusieurs groupes en fonction de la similitude des caractéristiques.

La dernière famille d'algorithmes de regroupement correspond aux cartes auto organisatrices permettant de réduire la dimensionnalité des données à des espaces de faible dimension. Les observations sont rassemblées dans des nœuds d'observations similaires. Les nœuds sont ensuite

répartis sur une carte où les nœuds similaires sont regroupés les uns à côté des autres [Hajjar, 2014].

- **Détection d'anomalie**

La détection d'anomalie est employée pour identifier les observations rares qui se distinguent de la majorité de l'ensemble de données. La machine analyse la structure des données et doit trouver les échantillons d'observations dont les caractéristiques s'écartent significativement de celles des autres échantillons. La détection d'anomalie est appliquée dans le développement des systèmes de sécurité, l'identification de menaces pour la cyber sécurité, la détection des fraudes bancaires, etc. Parmi les algorithmes de détection d'anomalie nous pouvons citer l'exemple qualifié de forêt d'isolation (*Isolation Forest*) qui permet d'isoler les observations atypiques [Liu *et al*., 2008] [Hmedna *et al*., 2020].

- **Réduction de dimension**

La réduction de dimension est essentielle lorsqu'il y a un nombre élevé de caractéristiques dans l'ensemble de données [Yang *et al*., 2018]. Cette dimensionnalité élevée augmente les risques d'*overfitting* et par conséquent diminue les performances du modèle. En effet, l'*overfitting* survient lorsque la machine œuvre bien avec les données d'apprentissage et œuvre mal avec de nouvelles données. Dans ce cas, le modèle est estimé incapable de généraliser. De plus, l'entraînement avec des données à haute dimensionnalité nécessite des ressources informatiques importantes en termes particulièrement de capacité de stockage et de puissance de calcul. Ces problèmes sont connus sous le nom de malédiction de la dimension ou celui de fléau de la dimension [Anowar *et al*., 2021]. Les techniques de réduction de dimension visent à surmonter les deux problèmes mentionnés. La machine analyse la structure des données et apprend comment la simplifier tout en conservant les principales informations. La réduction de dimension est souvent appliquée dans le but de simplifier la complexité des ensembles de données ce qui facilite l'apprentissage pour des problèmes de régression ou de classification. Parmi les algorithmes de réduction de dimension nous pouvons citer l'exemple de l'Analyse en Composantes Principales (*Principal Component Analysis*) [Alkhayrat *et al*., 2020] [Raj, 2021].

2.4.2 Evaluation de performance

Dans cette thèse nous traitons un problème de classification non supervisée (*clustering*/regroupement) pour analyser les expériences d'apprentissage. Nous exposons ci-dessous deux méthodes d'évaluation de performance du modèle de regroupement : indice de David Boudin et méthode silhouette que nous retiendrons pour évaluer l'une de nos deux contributions.

- **Indice de David Bouldin (DB)**

L'indice de David Bouldin est basé sur le principe des distances entre les clusters [Hidayat *et al.*, 2020]. Il cherche à mesurer à quel point un cluster est similaire au cluster qui lui est le plus proche. L'indice DB est formulé de la façon suivante [Davies et Bouldin, 1979] :

$$DB = \frac{1}{K}\sum_{k=1}^{K} \max_{j \neq i}\{D_{i,j}\}$$

Avec:

- K: le nombre de clusters,
- pour chaque cluster i, il faut trouver le cluster j qui maximise l'indice de similarité $D_{i,j}$ décrit comme suit :

$$D_{i,j} = \frac{d_i + d_j}{d_{ij}}$$

Où

- d_i représente la distance moyenne entre chaque point du i-ème cluster et le centroïde de ce cluster
- d_j représente la distance moyenne entre chaque point du j-ème cluster et le centroïde de ce cluster
- $d_{i,j}$ désigne la distance euclidienne entre les centroïdes des i-ème et j-ème clusters.

Il est généralement utilisé pour décider du nombre de clusters dans lesquels les points de données doivent être affectés. Cet indice vise à minimiser la distance moyenne entre chaque cluster et le cluster le plus similaire. Ainsi, la valeur minimale de l'indice David Bouldin indique le nombre optimal de clusters [Safaei-Farouji *et al.*, 2022]. En d'autres termes, la meilleure partition est celle qui minimise la similarité entre les clusters.

- **Méthode silhouette**

Le coefficient de silhouette moyen est utilisé pour évaluer la qualité du regroupement. Ce coefficient permet de savoir à quel point l'affectation de chaque observation à un cluster est correcte. Il est compris entre -1 et 1. Plus il tend vers 1 plus l'assignation d'une observation à son cluster est adéquate [Ansari *et al.*, 2015] [Safaei-Farouji *et al.*, 2022]. Il permet de savoir à quel point l'affectation de chaque apprenant à un cluster est correcte. Le coefficient de silhouette *s* de chaque apprenant *x* est donné par :

$$s(x) = \frac{b(x) - a(x)}{max\ (a(x), b(x)}$$

Avec :

- $a(x)$ est la distance moyenne de de l'apprenant x à tous les autres apprenants du cluster C_k auquel il appartient. Cette distance est calculée par :

$$a(x) = \frac{1}{|C_k\ | - 1} \sum_{u \in C_k\ , u \neq x} d(u, x)$$

- $b(x)$ est la plus petite valeur que pourrait prendre a(x), si l'apprenant x était assigné à un autre cluster. Cette valeur est calculée par :

$$b(x) = min_{l \neq k}\ \ \frac{1}{|C_l\ |} \sum_{u \in C_l} d(u, x)$$

Le coefficient de silhouette est compris entre -1 et 1. Plus il tend vers 1 plus l'assignation de l'apprenant à son cluster est adéquate.

2.4.3 Algorithmes incontournables de regroupement

Plusieurs algorithmes de regroupent sont disponibles. Nous en présentons quelques-uns usités et probants : Propagation d'affinité, Partitionnement spectral, Regroupement Aggloméré et K-moyenne.

- **Propagation d'affinité (*Affinity Propagation*)**

La propagation d'affinité est un algorithme itératif, reposant sur le partage des affinités, dont le principe est illustré dans la Figure 2.11. Il crée des clusters en envoyant des messages entre les points de données jusqu'à convergence. L'idée principale est de partir d'un graphe à travers lequel des messages sont transmis entre les données (les sommets du graphe) le long des arrêtes, en fonction de la similitude entre les données. L'algorithme assure la mise à jour de deux matrices :

la matrice de responsabilité et la matrice de disponibilité. En se basant sur la matrice de responsabilité, chaque point de données (observation) envoie des messages à tous les autres points les informant de son attractivité envers eux. La matrice de responsabilité *R* a des valeurs *r* (*i*, *k*) indiquant à quel point l'observation x_k est bien adaptée pour servir de représentant pour x_i, par rapport aux autres représentants candidats pour x_i. Chaque cible répond ensuite à tous les expéditeurs, en se référant à la matrice de disponibilité, en informant chaque expéditeur de sa disponibilité à s'associer avec lui, compte tenu des messages qu'elle a reçus de tous les autres expéditeurs. La matrice de disponibilité *A* contient des valeurs *a (i, k)* indiquant à quel point il serait approprié pour x_i de choisir x_k comme représentant, en tenant compte de la préférence des autres points. La procédure d'échange de messages se poursuit jusqu'à ce qu'un consensus soit atteint. Ces échanges doivent permettre de déterminer quelles données sont de bons représentants locaux et quels représentants modélisent le mieux chacune des autres données [Frey et Dueck, 2007]. Finalement, tous les points ayant le même représentant sont placés dans le même cluster. Contrairement aux autres algorithmes de regroupement, la propagation d'affinité ne nécessite pas de spécifier le nombre de groupes.

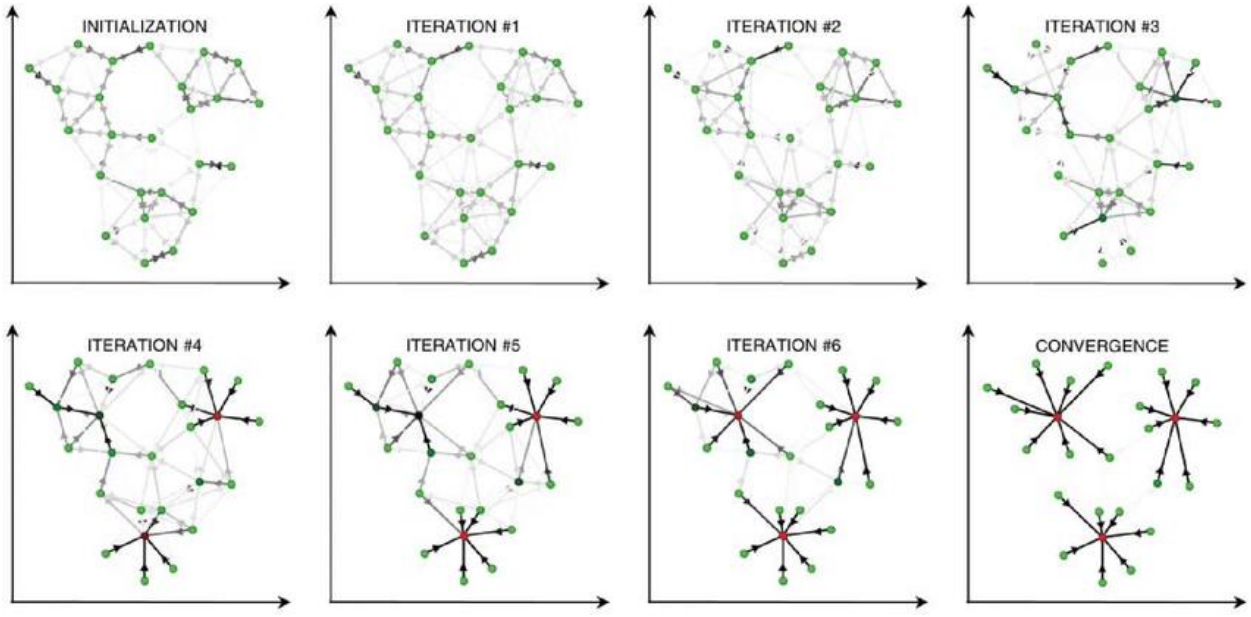

Figure 2.11 Exemple d'exécution de l'algorithme propagation d'affinité d'affinité [Torrent-Fontbona *et al*., 2012]

- **Partitionnement spectral (*Spectral Clustering*)**

Dans un partitionnement spectral, les observations sont représentées comme des nœuds d'un graphe connecté. Les clusters sont trouvés en partitionnant ce graphe, sur la base de sa décomposition spectrale, en sous-graphes. Les nœuds éloignés mais connectés appartiennent au même cluster et les points moins éloignés les uns des autres peuvent appartenir à des groupes différents s'ils ne sont pas connectés.

Comme le partitionnement spectral est basé sur la théorie des graphes, le jeu de données est représenté sous la forme d'une matrice de similarité entre toutes les observations. Soit un ensemble $E=\{x_i,\ 1\leq i \leq N\}$ de N observations décrites par la matrice de similarités S telle que l'élément $s_{ij} \leq 0$ correspond à la similarité entre x_i et x_j. L'objectif du partitionnement spectral est de répartir les N observations de l'ensemble E en k groupes disjoints $(E_1, \ldots, E_k)$ tels que les similarités soient fortes intra-groupe et faibles entre les groupes. Le partitionnement spectral procède en plusieurs étapes [Von Luxburg, 2007] :

1. Etape 1 : A partir de S un graphe de similarités G est déduit avec les pondérations tel que $w_{ij} = s_{ij}$. Dans ce graphe, chaque sommet correspond à une observation et chaque arête qui relie deux observations est pondérée par la similarité entre ces deux observations.
2. Etape 2 : A l'aide du graphe, une représentation vectorielle des observations est obtenue comme suit :
   - Calculer la matrice laplacienne L à partir de la matrice de degrés D et la matrice d'adjacence A en suivant la formule $L=D-A$
   - Extraire les k vecteurs propres $u_1, \ldots, u_k$ de L correspondant aux k plus petites valeurs de L,
   - Soit U_k la matrice dont les colonnes sont les vecteurs $u_1, \ldots, u_k$,
   - Soient $y_i \in \mathbb{R}^k$, $1 \leq i \leq N$, N nombre de lignes de la matrice U_k, chaque observation i est représentée par le vecteur y_i.
3. Etape 3 : Un algorithme de regroupement est enfin appliqué pour obtenir les k groupes :
   - Appliquer K-means par exemple aux N vecteurs y_i pour obtenir les groupes $C_1, \ldots, C_k$
   - Pour $1 \leq j \leq k$, $E_j=\{x_i | y_i \in C_j\}$.

- **Regroupement Aggloméré (*Agglomerative Clustering*)**

Soit $g(C_i, C_j)$ une fonction qui mesure la proximité entre les clusters C_i et C_j. Soit t le niveau actuel de la hiérarchie. Le schéma du regroupement aggloméré illustré dans la figure 2.12 peut alors s'énoncer comme suit [Ansari *et al.*, 2015] :

Initialisation :

Choisissez R_0 $\{Ci = \{x_i\}, i = 1, \ldots, N\}$ comme regroupement initial.

t=0.

Répéter :

$t = t+1$

Parmi toutes les paires de clusters (C_r, C_s) possibles en R_{t-1}, trouver celle (C_i, C_j) telle que

$$g(C_i, C_j) = \begin{cases} min_{r,s}\,(C_r, C_s) \text{ si g est une fonction de dissimilarité} \\ max_{r,s}\,(C_r, C_s) \text{ si g est une fonction de similarité} \end{cases}$$

Définir $C_q = C_i \cup C_j$ et produire le nouveau regroupement $R_t = (R_{t-1} - \{C_i, C_j\}) \cup \{C_q\}$.

Jusqu'à

ce que tous les vecteurs se trouvent dans un seul cluster.

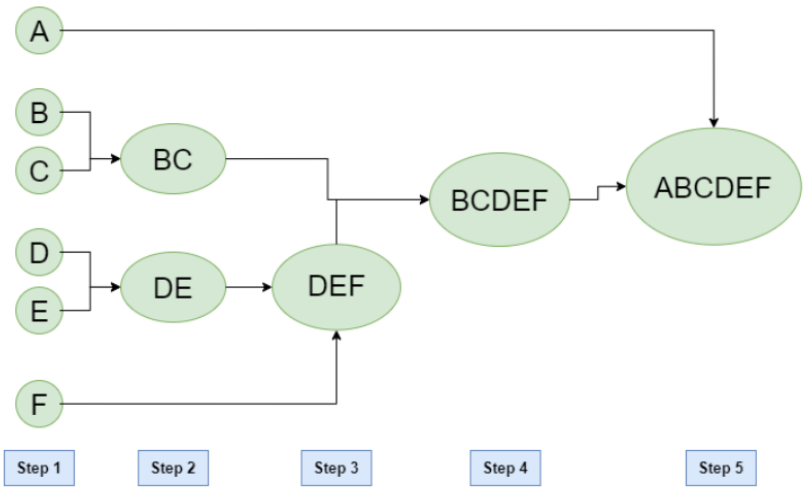

Figure 2.12 Exemple d'exécution de l'algorithme de regroupement aggloméré [Nawrin *et al.*, 2017]

- **K-moyenne (*K-means*)**

L'algorithme k-means consiste à regrouper des observations en k groupes de façon à minimiser la distance au carré entre le centroïde d'un cluster et chaque observation appartenant au cluster.

La philosophie de k-means consiste à associer chaque apprenant au cluster le plus proche de façon à minimiser la variation intra-cluster (*within-cluster variation*) qui correspond à la somme des distances euclidiennes au carré entre chaque apprenant et le centroïde de son cluster [Tao *et al.*, 2016]. Une des difficultés de mise en œuvre de k-means est le choix du nombre de clusters k qui doit être spécifié au préalable. Un meilleur choix de *k* mène à un meilleur regroupement qui minimise les distances intra-clusters et maximise les distances inter-clusters. Il existe des méthodes pour déterminer le nombre optimal de clusters *k* comme la méthode Elbow [Madhulatha, 2012] et la méthode silhouette [Ansari *et al.*, 2015] [Mosavi et Safaei-Farouji, 2021]. Il convient ici de décrire le fonctionnement de l'algorithme k-means.

Entrée :

k : le nombre de clusters sélectionné

X (n, p) : la matrice de données relative aux n apprenants.

Début

Choisir aléatoirement *k* apprenants pour initialiser les centres de gravité des clusters nommé aussi centroïdes des clusters.

REPETER

Calculer la distance euclidienne entre les apprenants de la matrice de donnée *X (n, p)* et chaque centroïde

Affecter chaque apprenant au cluster dont il est le plus proche à son centroïde

Recalculer les centres de gravité de chaque cluster qui deviennent les nouveaux centroïdes

JUSQU'A stabilisation des centres de clusters

Fin

Dans le contexte du e-learning, Moubayed et al., ont eu recours à cet algorithme pour identifier le niveau d'engagement des apprenants [Moubayed *et al.*, 2020]. Ce travail de recherche a été détaillé dans la section 1.3.1.

Afin de regrouper les apprenants selon leur expérience d'apprentissage nous avons eu recours à cet algorithme. Dans le chapitre 3 nous précisons les critères de choix de cet algorithme. Dans le chapitre 4, k-means a été concrétisé sur deux cas d'étude réels ; les résultats des expérimentations y sont présentés et analysés.

2.5 Discussion

Dans le cadre de cette thèse, nous visons à aider les concepteurs pédagogiques à évaluer automatiquement leurs contenus éducatifs en ligne afin de pouvoir les améliorer. Conventionnellement, l'évaluation des contenus éducatifs en ligne est basée sur l'opinion des apprenants et des autres parties prenantes clés comme les parents et les experts pédagogiques [Margaryan *et al.*, 2015]. Le questionnement des apprenants à propos de leur perception permet de recueillir des informations utiles que les prestataires de formations en ligne pourront exploiter. Toutefois, ces informations peuvent, facilement, être faussées ou bruitées dès que les apprenants répondent de manière non réfléchie ou font des erreurs de diagnostic [Melesko et Kurilovas, 2018]. C'est la raison pour laquelle certains chercheurs comme Melesko et Kurilovas considèrent que les apprenants n'ont pas l'expertise nécessaire pour évaluer la conception pédagogique [Melesko et Kurilovas, 2018]. Aussi, il est important de prendre en considération les connaissances des experts pédagogiques pour évaluer les contenus éducatifs en ligne [Margaryan *et al.*, 2015].

Quatre critères justifient notre intérêt pour l'apprentissage automatique au service de l'évaluation des contenus éducatifs en ligne.

Le premier critère est sa capacité, son aptitude à analyser un grand volume de données en un temps réduit [Handelman *et al.*, 2018] [Shang et You, 2019]. Dans la pratique, la mission de l'expert pédagogique est ardue. Celui-ci doit d'abord observer le comportement de plusieurs apprenants à travers leurs traces numériques d'interaction pour pouvoir analyser leurs expériences d'apprentissage et les interpréter et ainsi prendre les décisions éclairées, évaluer le contenu éducatif. Généralement, les traces des apprenants sont recueillies à partir de l'EIAH ou à partir d'un navigateur web. A l'état brut, ces traces prennent des formes diversifiées [Samuelsen *et al.*, 2019]: représentation graphique [Jivet *et al.*, 2018], fichier texte [Wong *et al.*, 2018], fichier log [ElSayed *et al.*, 2019], fichier csv [Pardos et Kao, 2015], etc. Sans l'apprentissage automatique, il est maintes fois fastidieux et infructueux de visualiser et d'analyser cette quantité énorme de traces décrivant toutes les actions des apprenants.

Le second critère réside dans l'objectivité de l'analyse [Fatima and Pasha, 2017]. Celle-ci étant automatique, un ensemble d'erreurs humaines, liées à l'observation, l'analyse et l'interprétation des expériences d'apprentissage peuvent être évitées.
Le troisième critère consiste dans la facilité d'implémentation en comparaison avec la programmation qualifiable de traditionnelle [Chollet, 2017] [Geisslinger, 2019]. L'automatisation du processus d'évaluation des contenus éducatifs peut aider les prestataires de formations en ligne à réduire les tâches manuelles de l'expert et obtenir de meilleurs résultats. L'exercice sera insoutenable sinon impossible avec la programmation qualifiable de traditionnelle parce qu'il faudrait définir toutes les tâches devant être effectuées de la même manière que l'expert. Par conséquent, le programmeur sera face à un grand nombre de cas possibles à coder. Avec l'aide de l'apprentissage automatique, ce problème pourrait être résolu. La mission pourra être ainsi gérée plus rapidement, ce qui devrait favoriser un gain de terrain sur les prestataires concurrents.

Le quatrième et dernier critère est la disponibilité d'outils d'évaluation de performance des solutions s'appuyant sur l'apprentissage automatique [Davies et Bouldin, 1979] [Ansari *et al.*, 2015] [Refaeilzadeh and Liu, 2009] [Boutaba *et al.*, 2018] [Hussain *et al.*, 2019] [Juba et Le, 2019] [Luque *et al.*, 2019] [Padilla *et al.*, 2020] [Lopez-Bernal *et al.*, 2021] [Safaei-Farouji *et al.*, 2022].
Aussi, sur la base de ces quatre critères, nous nous proposons de recourir à l'apprentissage automatique pour aider le concepteur pédagogique à prendre une décision éclairée et objective, pilotée par les données des apprenants.

2.6 Conclusion

Dans ce chapitre, nous avons passé en revue les différentes méthodes existantes pour l'apprentissage automatique. Nous avons, particulièrement, mis la lumière sur la classification non supervisée (le regroupement/*clustering*) et la classification supervisée. Nous avons présenté, pour la classification supervisée, les algorithmes : perceptron, adaline dans ses deux versions avec la descente de gradient et la descente de gradient stochastique, la machine à vecteur de support et la régression logistique. Nous avons également exploré les différentes mesures qui peuvent être utilisées pour évaluer la qualité de la performance d'un modèle de classification. Pour la classification non supervisée/le regroupement nous avons évoqué les méthodes hiérarchiques et celles de partitionnements. Nous avons décrit les algorithmes de propagation d'affinité, de

regroupement spectral, de regroupement aggloméré et k-moyennes avec les différents indices existant dans la littérature pour évaluer le nombre optimal de clusters et la qualité du regroupement. Tous ces fondements théoriques relatifs à l'apprentissage automatique que nous avons sélectionnés ont permis de mettre en place une méthodologie spécifique à notre problématique d'évaluation automatique des contenus éducatifs en ligne, décrite dans le chapitre suivant.

Dans ce travail de recherche, nous projetons de donner à la machine la capacité d'apprendre toute seule à exécuter les tâches de l'expert pédagogique : l'évaluation des contenus éducatifs en ligne. A cet égard, nous proposons deux contributions :

- Une approche d'analyse multicritère des expériences d'apprentissage dans un EIAH (MALEA : *Multicriteria Approach for Learning Experience Analysis*), basée sur l'algorithme d'apprentissage automatique non supervisé k-moyenne (*k-means*), Cette approche est capable de regrouper les apprenants selon leur comportement. MALEA vise à permettre, plus précisément, d'identifier automatiquement les groupes d'apprenants ayant des expériences d'apprentissage similaires.
- Une approche de prédiction de la réussite des contenus éducatifs en ligne et de recommandation en vue de leur amélioration (ACSP : *Approach for Content Success Prediction*). Basée sur l'apprentissage automatique supervisé, cette approche permet une classification binaire des contenus éducatifs en ligne. Pour ce faire, nous recourrons à la régression logistique.

Chapitre 3 : Système d'aide à l'évaluation et l'amélioration des contenus éducatifs en ligne

3.1 Introduction

L'évaluation des contenus éducatifs en ligne par l'expert pédagogique, tâche complexe, empreinte de subjectivité et coûteuse en termes de temps, nous amène à réfléchir à un procédé automatisé, objectif et rentable. Ce chapitre présente la méthodologie générale proposée pour aider concepteurs et expert pédagogiques à évaluer et ainsi améliorer leurs contenus éducatifs en ligne partant d'une analyse multicritère des expériences d'apprentissage. Dans la première section de ce chapitre, nous présentons le principe général de notre système d'aide à l'évaluation et l'amélioration des contenus éducatifs en ligne couplant nos deux contributions :

- Une approche d'analyse multicritère des expériences d'apprentissage (MALEA : *Multicriteria Approach for Learning Experience Analysis*) basée sur le regroupement des apprenants à l'aide de la méthode k-means.

- Une approche de prédiction de la réussite des contenus éducatifs (ACSP : *Approach for Content Success Prediction*) basée sur la première approche proposée MALEA et la méthode de régression logistique.

Dans la deuxième section, nous présentons, d'abord, notre première approche MALEA et nous justifions le choix d'intégrer la méthode k-means dans cette approche. Ensuite, nous détaillons notre approche MALEA et son utilisation. Dans la troisième section, nous présentons notre approche ACSP. Nous décrivons la méthode de la régression logistique et nous expliquons sa combinaison avec MALEA dans l'approche ACSP.

3.2 Contributions méthodologiques

3.2.1 Principe général des contributions

Dans le cadre de cette thèse, nous proposons un système intelligent d'aide à la décision qui permet d'assister les concepteurs pédagogiques dans leurs tâches d'évaluation et d'amélioration de contenu éducatif en ligne. Ce système est générique. Il peut être adaptable pour évaluer différents contenus en ligne, à savoir les contenus publiés sur les sites de e-commerce, sur les réseaux sociaux, etc., dans le but de les améliorer. Notre système intelligent d'aide à la décision pédagogique combine les deux approches proposées :

1. **Une approche d'analyse multicritère des expériences d'apprentissage (MALEA : *Multicriteria Approach for Learning Experience Analysis*) :**

   Cette approche se base sur la méthode d'apprentissage automatique non supervisé k-means [Ostrovsky *et al*., 2013] [Baruri *et al*., 2019]. Son rôle est de regrouper les apprenants ayant des expériences d'apprentissage similaires dans un même cluster. L'apport majeur de cette approche comme présenté et discuté dans les sections 1.3.2 et 0 du premier chapitre réside dans sa capacité à prendre en compte plusieurs critères dans le processus d'analyse des expériences d'apprentissage. MALEA constituera le socle de l'évaluation des contenus éducatifs en ligne. Il s'agit d'étiqueter/labéliser les contenus éducatifs par le label « réussi » ou « à améliorer ».

   Dans le Tableau 3.1 ci-dessous nous comparons MALEA aux méthodes existantes discutées dans la section 1.3 du premier chapitre.

Tableau 3.1 Comparaison entre MALEA et les méthodes évoquées dans l'état de l'art

Méthodes existantes	**MALEA**
Observation, analyse et interprétation humaine des expériences d'apprentissage	Analyse automatique des expériences d'apprentissages
Méthodes empiriquement menées à travers des enquêtes de satisfaction.	Utilise les données des apprenants générées automatiquement dans les EIAH
Subjectives	Objective
Analyse suivant un seul ou deux critères	Analyse multicritère

2. Une approche de prédiction de la réussite des contenus éducatifs en ligne (ACSP : *Approach for Content Success Prediction*) :

Cette approche permet au concepteur pédagogique d'évaluer son contenu éducatif à n'importe quel stade de son élaboration et notamment avant sa diffusion sur un Environnement Informatique pour l'Apprentissage Humain (EIAH). Combinant la régression logistique et MALEA, ACSP permet de se prémunir contre l'imprécision éventuelle du jugement humain affectant le processus de décision.

Avant de détailler les approches MALEA et ACSP, respectivement, dans les sections 3.3 et 3.4, nous expliquons comment elles sont combinées et employées dans notre système d'aide à l'évaluation et l'amélioration des contenus éducatifs en ligne. La Figure 3.1 illustre l'architecture générale de ce système intelligent d'aide à la décision pédagogique qui pourrait être intégré sous forme de module dans un Environnement Informatique pour l'Apprentissage Humain (EIAH).

En effet, notre système d'aide à la décision est fondé sur la technologie de l'apprentissage automatique. Rappelons que le principe de l'apprentissage automatique consiste à donner à la machine, sans la programmer de façon explicite, la capacité d'apprendre à exécuter une tâche, à partir d'expériences, c'est-à-dire d'exemples de données. L'une des stratégies de l'apprentissage automatique est l'apprentissage automatique supervisé. Celui-ci consiste à fournir à la machine un jeu de données labellisées (X, Y) et de lui demander de trouver la fonction d'association entre les variables prédictives en entrée X et la variable à prédire Y appelée aussi label. L'objectif est de rendre la machine capable de prédire le label inconnu y associé à une nouvelle observation/donnée x.

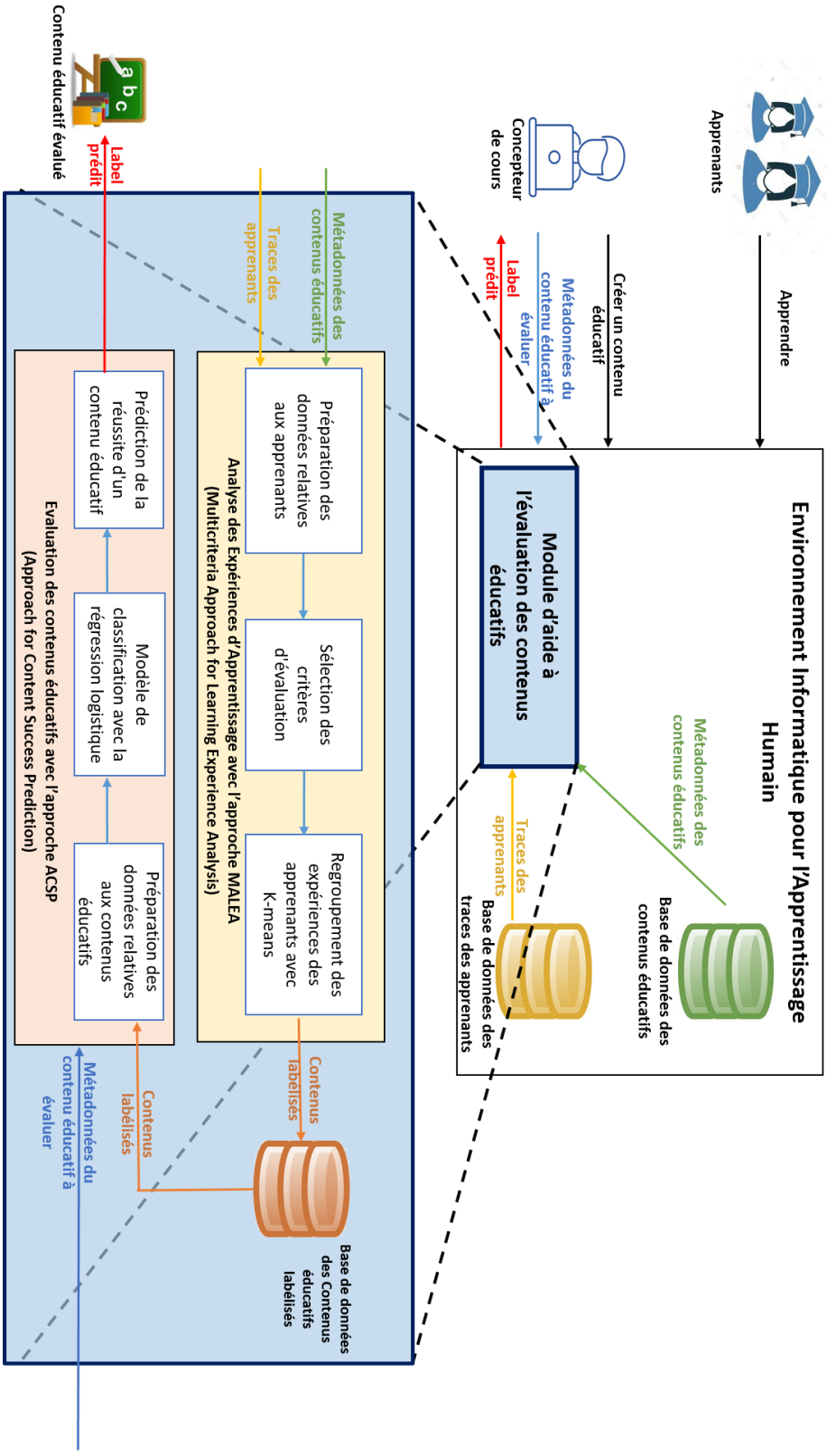

Figure 3.1 Architecture générale du système d'aide à l'évaluation et l'amélioration des contenus éducatifs en ligne

Dans notre cas d'étude, nous aspirons à construire un modèle prédictif capable de distinguer un contenu éducatif réussi d'un contenu éducatif à améliorer. Nous considérons cette tâche comme une tâche de classification, un type de problème résoluble par l'apprentissage automatique supervisé. S'agissant d'un problème d'apprentissage automatique supervisé, un jeu de données labellisées (*X*, *Y*) devrait être disponible. Ainsi, la classification est précédée d'une phase de labellisation dans laquelle, à chaque contenu éducatif *x*, est associé un label *y* indiquant sa classe.

Comme cela, nous disposons d'une base de données de contenus éducatifs en ligne. Les données d'interactions des apprenants suivant ces contenus éducatifs sont collectées et stockées dans la base de données des traces des apprenants. Pour chaque contenu éducatif, les expériences d'apprentissages sont analysées avec l'approche MALEA se basant sur les traces des apprenants. Le décideur pédagogique, en connaissance des résultats des analyses produites par MALEA est en mesure d'évaluer le contenu éducatif en question. Toute évaluation sera stockée dans la base de données des contenus éducatifs labélisés. A partir de cette base de données, le modèle de classification des contenus éducatifs en ligne est construit avec l'approche ACSP permettant au décideur pédagogique, qu'il soit concepteur de contenu ou expert d'évaluation de contenu, d'évaluer afin d'améliorer un contenu éducatif en ligne à n'importe quel stade de son élaboration et notamment avant sa diffusion sur l'Environnement Informatique pour l'Apprentissage Humain (EIAH).

3.3 Approche d'analyse multicritère des expériences d'apprentissage (MALEA)

Dans cette section nous présentons notre approche d'analyse multicritère des expériences d'apprentissage MALEA en spécifiant le rôle qu'elle joue dans le système d'aide à l'évaluation et l'amélioration des contenus éducatifs en ligne. Nous commençons par la définition de la problématique qui sera traitée par cette approche. Puis, nous détaillons les différentes étapes de MALEA tout en justifiant l'utilisation de l'algorithme de regroupement k-means.

3.3.1 Enoncé du problème traité par l'approche MALEA

Afin d'évaluer un contenu éducatif en ligne, nous considérons qu'il est nécessaire d'analyser au préalable les expériences d'apprentissage relatives aux apprenants ayant suivi ce contenu. Pour

ceci, nous proposons l'approche MALEA. La Figure 3.2 schématise la résolution du problème d'évaluation des contenus éducatifs à l'aide de notre approche.

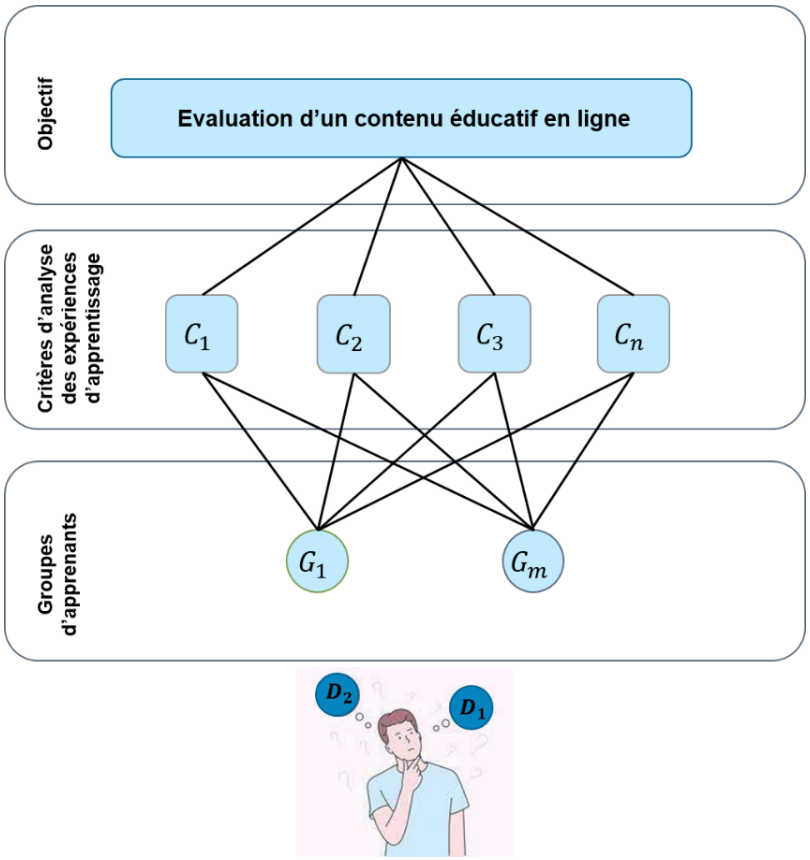

Figure 3.2 Formulation du problème d'évaluation du contenu éducatif en ligne traité par MALEA

Soit un ensemble d'apprenants X et n critères d'analyse des expériences d'apprentissage $\boldsymbol{C_i}$ où i = (1, ..., n). Le problème consiste à regrouper l'ensemble X des apprenants, suivant un même contenu éducatif en ligne, dans m clusters homogènes $\boldsymbol{G_j}$ où j = (1, ..., m) tout en tenant compte des n critères sélectionnés par le/les décideurs pédagogiques. Le résultat du regroupement va permettre aux décideurs impliqués de comprendre l'impact du contenu éducatif sur les expériences d'apprentissage des apprenants du point de vue des n critères. Ceci éclaire et appuie le décideur dans l'évaluation du contenu éducatif en ligne et la prise de décision soit ($\boldsymbol{D_1}$) contenu éducatif réussi ou ($\boldsymbol{D_2}$) contenu éducatif à améliorer.

3.3.2 Choix de l'algorithme k-means

Afin d'identifier les groupes d'apprenants ayant des expériences d'apprentissage similaires, nous avons sélectionné l'algorithme de *clustering* k-means. Ce choix est fait sur la base de plusieurs critères.

D'abord, k-means est l'un des algorithmes couramment utilisés pour résoudre les problèmes de partitionnement dans un ensemble de données non labélisées [Wu, 2021] [Alzubi *et al*., 2018] ; tel est notre cas.

En outre, il a montré une efficacité considérable [Yue *et al*., 2018] [Sinaga et Yang, 2020]. Aujourd'hui, les champs d'application de k-means sont divers. Il est employé dans la segmentation des clientèles, la classification des documents en fonction de leurs contenus, la segmentation et la compression d'images, etc. [Lithio et Maitra, 2018].

K-means est caractérisé aussi par sa vitesse rapide d'apprentissage et son faible coût de calcul [Tang *et al*., 2017].

Par ailleurs, l'un des avantages de k-means réside dans la simplicité de sa mise en œuvre et la clarté avec laquelle il présente les résultats, ce qui maximise la compréhension du regroupement.

Dans le Tableau 3.2 nous comparons k-means avec les méthodes de *clustering* : k-means, regroupement aggloméré, propagation d'affinité et partitionnement spectral.

Tableau 3.2 Comparaison entre les méthodes de *clustering* : k-means, regroupement aggloméré, propagation d'affinité et partitionnement spectral [scikit-learn]

Algorithme	**K-means**	**Regroupement aggloméré (Hiérarchique ascendant)**	**Propagation d'affinité**	**Partitionnement spectral**
Paramètres	Il nécessite une connaissance préalable du nombre de clusters que l'on souhaite avoir	Nous pouvons nous arrêter à n'importe quel nombre de clusters que l'on trouve approprié	Elle nécessite le choix des deux paramètres : préférence et facteur d'amortissement. La préférence correspond aux observations pouvant être des exemplaires (clusters). L'estimation du nombre de clusters dépend de la préférence.	Il nécessite une connaissance préalable du nombre de clusters que l'on souhaite avoir

			Le facteur d'amortissement amortit la responsabilité et la disponibilité des messages pour éviter les oscillations numériques lors de la mise à jour des messages.	
Forme des clusters	Regroupement en clusters séparés ayant généralement une forme sphérique ou spectrale	Regroupement en un ensemble de clusters imbriqués qui sont disposés comme un arbre. Cet algorithme est particulièrement utile lorsque l'objectif est d'organiser les clusters dans une hiérarchie.	Regroupement en clusters séparés	Regroupement en clusters ayant une forme spectrale
Géométrie	Distance entre les points	Distance par pair	Graphe des plus proches voisins	Graphe des plus proches voisins
Calcul	Calcul simple	Il nécessite le calcul et le stockage d'une matrice de distance n×n. Pour les très grands ensembles de données, cela peut être coûteux et lent.	Méthode complexe	Il peut être lent pour les graphes hautement connectés
Cas d'usage	Fonctionne avec des attributs numériques Usage général, clusters de taille similaire,	Fonctionne avec tous les types d'attributs Nombreux clusters avec d'éventuelles contraintes de connectivité,	Nombreux clusters, taille des clusters asymétrique, géométrie non plate, inductive.	Quelques clusters de taille similaire, géométrie non plate, transductive

	géométrie plate, inductive	distance non euclidienne, transductive		
Scalabilité	Grand nombre d'observations, nombre de clusters moyen.	Grand nombre d'observations et de clusters	Méthode complexe ne supportant pas un grand nombre d'observations.	Nombre d'échantillons moyen, faible nombre de clusters

3.3.3 Démarche de l'approche MALEA

MALEA est l'approche d'analyse des expériences d'apprentissage proposée dans le cadre de cette thèse pour évaluer les contenus éducatifs en ligne. Cette approche se caractérise par sa généricité et sa capacité d'être configurée pour résoudre différents problèmes d'analyse d'expérience utilisateur.

MALEA repose sur cinq étapes comme le montre la Figure 3.3 ci-dessous, notamment :

- Étape 1 : Collecte des données des apprenants.
- Étape 2 : Préparation des données.
- Étape 3 : Regroupement des apprenants avec l'algorithme k-means.
- Étape 4 : Identification des clusters.
- Étape 5 : Evaluation du contenu éducatif.

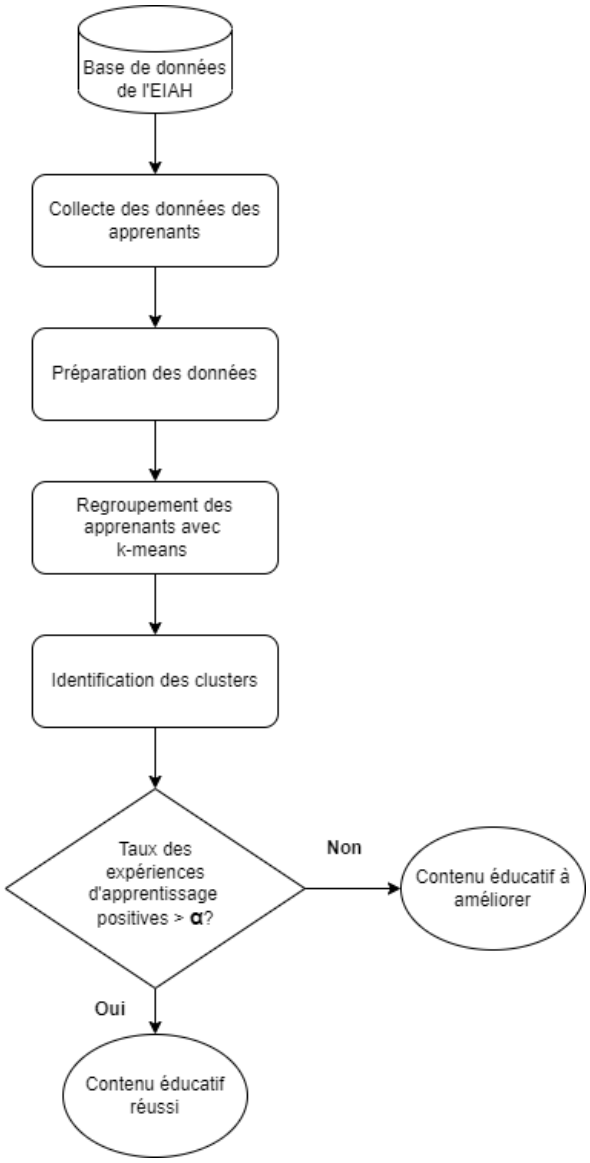

Figure 3.3 Processus de l'approche MALEA

Étape 1. Collecte des données des apprenants : Tout EIAH génère automatiquement des données relatives aux contenus éducatifs en ligne, aux apprenants et à leurs interactions [Mourali *et al.*, 2021a]. Au cours de cette étape nous collectons les traces numériques d'interaction d'un ensemble d'apprenants suivant un même contenu éducatif (nombre d'accès au cours, nombre de publications dans les forums de discussion, moyenne des notes obtenues dans les évaluations, etc.). En général, ces données sont utilisées pour comprendre le comportement des apprenants, afin d'améliorer l'enseignement et l'apprentissage en ligne [Herodotou *et al.*, 2019]. Nous exploitons ces données particulièrement à des fins d'amélioration des contenus éducatifs en ligne. Les

données brutes recueillies à partir des EIAH sont hétérogènes [Abdelouarit *et al.*, 2020] et souvent présentes sous forme de fichiers csv, fichiers journaux, etc.

Étape 2. Préparation des données : Cette étape concerne les opérations qui doivent être appliquées aux données brutes avant leur traitement et analyse. Comme MALEA est une approche d'analyse multicritère, c'est au cours de cette étape qu'il sera question de sélectionner les critères permettant d'analyser les expériences d'apprentissage. En effet, chaque critère permet d'évaluer un aspect spécifique de l'expérience d'apprentissage comme l'engagement de l'apprenant, sa performance aux tests et examens, son taux de satisfaction et son achèvement. Au cours de cette étape, les données apportant toute information utile sur les critères sélectionnés sont extraites, structurées, nettoyées et formatées. Les données finales sont représentées par la matrice X (n, p) suivante :

$$X(n,p) = \begin{pmatrix} x_1^1 & \dots & x_p^1 \\ & & \\ x_1^n & \dots & x_p^n \end{pmatrix}$$

La matrice X (n, p) considère n apprenants où chaque apprenant est représenté par un vecteur décrivant son expérience d'apprentissage. Chaque vecteur est formé par p caractéristiques. Ces caractéristiques traduisent les critères d'analyse des expériences d'apprentissage sélectionnés. Un critère peut être traduit (référé) par une caractéristique ou plus.

Étape 3. Regroupement des apprenants avec l'algorithme k-means : Pour analyser les expériences d'apprentissage, nous avons eu recours à un apprentissage automatique non supervisé. Nous avons opté pour un *clustering* (regroupement). C'est l'une des techniques d'analyse exploratoire des données capable de faire des inférences à partir de jeux de données en se servant uniquement des vecteurs d'entrée sans faire référence à des résultats connus ou labélisés. Cette technique est utilisée, particulièrement, pour identifier des sous-groupes homogènes dans un ensemble de données.

Afin de regrouper les apprenants, l'algorithme k-means est basé sur l'examen de la similarité entre les apprenants. La similarité entre deux apprenants X_i $(x_i^1, x_i^2, \dots, x_i^p)$ et X_j $(x_j^1, x_j^2, \dots, x_j^p)$ de la matrice X (n, p) est inférée grâce au calcul de la distance euclidienne d séparant leurs caractéristiques [Jiang *et al.*, 2020].

$$d(X_i, X_j) = \sqrt{\left(x_i^1 - x_j^1\right)^2 + \left(x_i^2 - x_j^2\right)^2 + \dots + \left(x_i^p - x_j^p\right)^2}$$

Nous utilisons le coefficient de silhouette moyen ainsi que l'indice de Davies Bouldin pour évaluer la qualité du regroupement. D'une part, car le coefficient de silhouette moyen nous permet d'évaluer l'homogénéité des clusters en vérifiant à quel point l'affectation de chaque apprenant à un cluster est correcte [Ghribi *et al.*, 2010]. D'autre part, car l'indice de Davies Bouldin indique à quel point les clusters sont séparés [Ghribi *et al.*, 2010]. Le Tableau 3.3 compare ces deux méthodes d'évaluation.

Tableau 3.3 Comparaison entre les deux méthodes d'évaluation de la qualité du regroupement : Silhouette et Davies Bouldin

Silhouette	**Davies Bouldin**
Il traite chaque observation en particulier.	Il traite chaque cluster individuellement.
Il permet de vérifier si chaque observation a été bien classé.	Il cherche à mesurer à quel point un cluster est similaire au cluster qui lui est le plus proche.
La meilleure partition des données est celle qui maximise la similarité entre les observations d'un même cluster.	La meilleure partition des données est celle qui minimise la similarité entre les clusters.
Evaluation de l'homogénéité des clusters.	Evaluation de la séparation entre les clusters.

Étape 4. Identification des clusters : Dans cette étape nous interprétons le résultat du regroupement fourni par l'algorithme k-means. Cette étape désigne, d'une part, l'identification des clusters. D'autre part, elle nous permet de faire ressortir les patterns cachés dans notre jeu de données en regroupant les apprenants qui se ressemblent. Dans notre cas d'étude, nous avons intérêt à regrouper les apprenants en deux groupes : ceux ayant des expériences d'apprentissage positives et d'autres ayant des expériences d'apprentissage négatives.

Étape 5. Evaluer le contenu éducatif : Dans cette étape, il est utile de déterminer la taille de chacun des clusters obtenus. Selon la taille de ces deux groupes d'apprenants, nous pouvons décider si le contenu éducatif nécessite une amélioration ou non. L'approche MALEA est basée sur un seuil d'évaluation (α) fixé par l'équipe pédagogique selon ses objectifs. Si la taille du cluster relatif aux apprenants ayant des expériences d'apprentissage positive dépasse le seuil (α), le contenu éducatif est considérée réussi. Au cas où la taille de ce cluster est inférieure au seuil (α), le contenu éducatif est considéré à améliorer. En se réfèrent à l'échelle de Likert [Joshi *et al.*, 2015], l'équipe pédagogique pourrait, par exemple, se fixer un seuil d'évaluation égal à 60% du

nombre d'apprenants ayant eu des expériences positives pour considérer le contenu éducatif comme réussi.

Dans ce qui suit, nous expliquons comment prédire la réussite d'un contenu éducatif en ligne avant ou après sa diffusion sur l'EIAH, à l'aide de l'approche ACSP que nous proposons.

3.4 Approche de prédiction de la réussite des contenus éducatifs en ligne (ACSP)

Dans cette section nous présentons notre approche ACSP de prédiction de la réussite des contenus éducatifs en ligne. ACSP est fondée sur notre approche d'analyse des expériences d'apprentissage MALEA et la technologie de l'apprentissage automatique supervisé, qui consiste en une classification binaire avec régression logistique [Ray, 2019] [Kohli *et al*., 2021] [Catal *et al*., 2019 (cf. section 2.3 du chapitre 2). Nous définissons, d'abord, les différentes notions du problème traité par ACSP. Ensuite, nous décrivons la démarche de cette approche et son architecture.

3.4.1 Formulation du problème traité par ACSP

Notre problème consiste à prédire si le contenu éducatif créé par le concepteur pédagogique est réussi, ou bien s'il nécessite des améliorations pour satisfaire les besoins de l'apprentissage. Il y a donc deux options possibles nous permettant de classer notre prédiction dans l'une ou l'autre catégorie. Il s'agit bien d'un problème de classification binaire que nous résolvons avec notre approche ACSP en utilisant la régression logistique, tel que présenté dans la Figure 3.4 ci-dessous. Rappelons que l'objectif de la classification supervisée, tâche de l'apprentissage automatique supervisé, consiste à entraîner la machine à classifier les contenus éducatifs d'une manière efficace. L'apprentissage automatique supervisé comporte principalement deux phases : apprentissage puis prédiction. Au cours de la phase d'apprentissage dite aussi d'entrainement, la machine ou l'algorithme de classification sélectionné a besoin d'exemples de données pour construire son système de raisonnement (modèle). Ainsi, nous fournissons un jeu de données d'entrainement contenant deux types de variables :

- La variable objectif ou label *Y*, c'est ce qu'on veut que la machine apprenne à prédire. Dans notre cas, c'est la classe qui permet d'identifier si le contenu éducatif est réussi ou s'il nécessite des améliorations.

- Les caractéristiques décrivant les contenus éducatifs *C* et qui influencent la valeur de la variable prédite $\hat{Y}$.

Ce jeu de données permet la construction du modèle *F* qui effectue la tâche de prédiction de la réussite de nouveaux contenus éducatifs dans la deuxième phase. Le modèle prédictif est défini par :

$$Y = F(C)$$

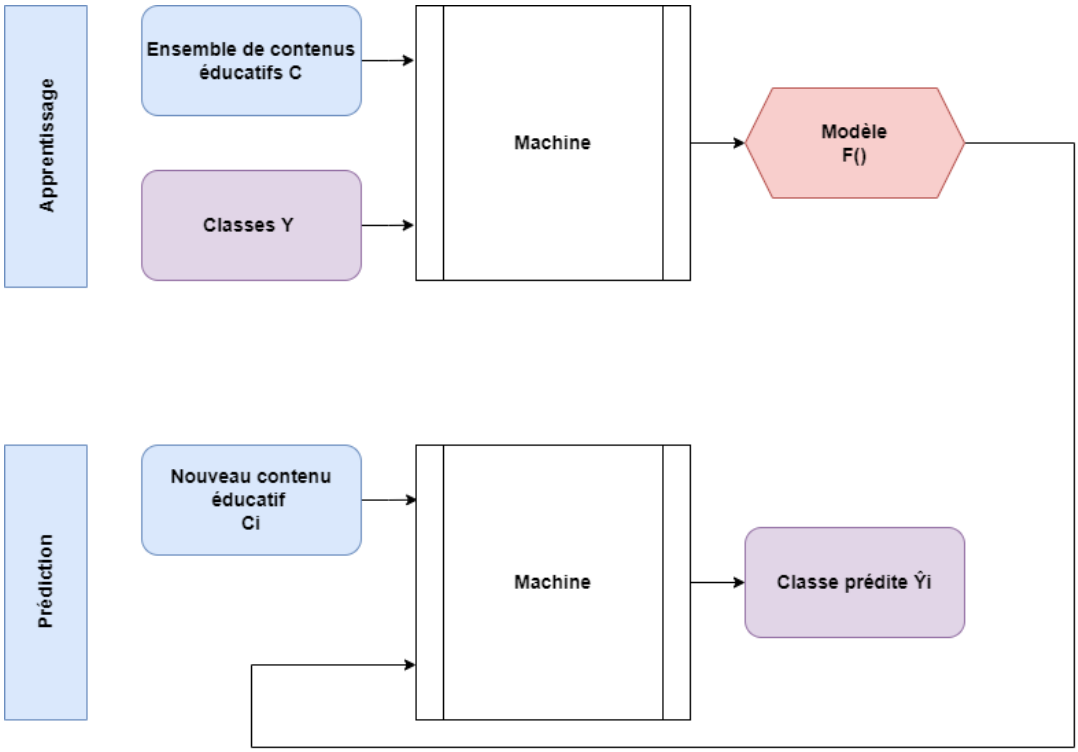

Figure 3.4 Formulation du problème de prédiction de la réussite des contenus éducatifs en ligne

3.4.2 Choix de la régression logistique

Plusieurs raisons nous motivent et orientent vers la sélection de l'algorithme de la régression logistique. D'abord, la régression logistique fait partie des algorithmes d'apprentissage supervisé. C'est une technique appropriée pour résoudre le problème de classification binaire où la variable label Y est une variable discrète [Trehan et Joshi, 2018] ; tel est notre cas (y_i =0 si le contenu éducatif est à améliorer ou y_i =1 si le contenu éducatif est réussi). De plus, c'est un algorithme largement cité dans l'état de l'art et couramment employé pour sa robustesse dans le développement des modèles analytiques [Peng *et al.*, 2002] [Alzubi *et al.*, 2018] [Ray, 2019]. Il est d'ailleurs au sommet des algorithmes les plus utilisés pour la prédiction dans le domaine du e-learning [Dalipi *et al.*, 2018] tel que montré dans la Figure 3.5.

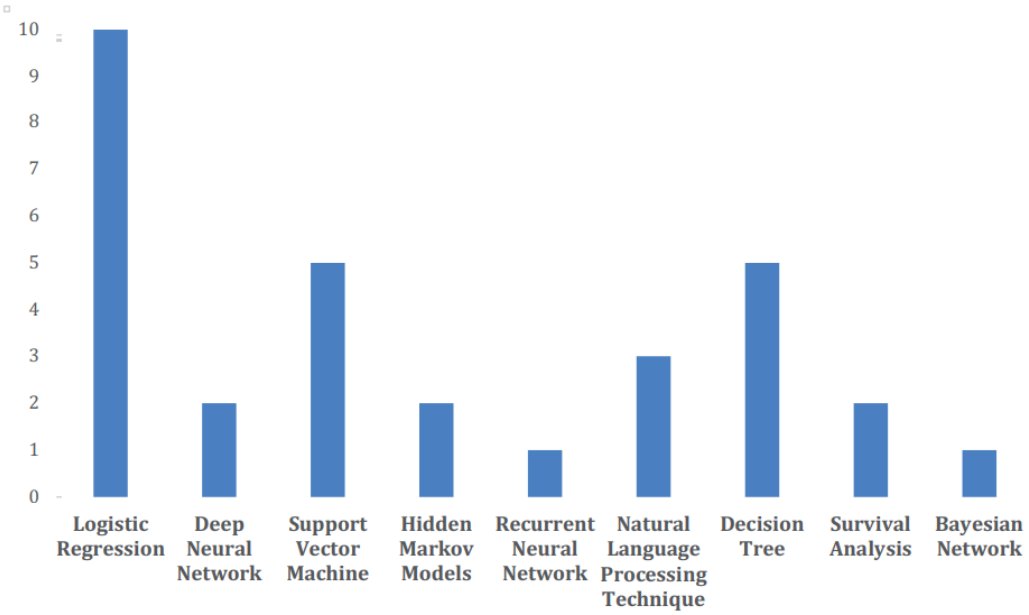

Figure 3.5 Fréquence de l'utilisation de la régression logistique dans la prédiction de l'abandon dans le domaine du e-learning [Dalipi *et al.*, 2018]

En outre, la régression logistique ne permet pas seulement de classifier un élément, mais précise également la probabilité de son appartenance à la classe où il a été associé [Lever et Altman, 2016]. Elle pourrait dans notre cas nous indiquer, par exemple, la probabilité qu'un contenu éducatif soit réussi. Ceci présente un avantage important par rapport aux autres algorithmes de classification non probabilistes tels SVM (*Support Vector Machines*) [Chauhan *et al.*, 2019], KNN (*K-Nearest Neighbors*) [Beckmann *et al.*, 2015], RNN (*Recurrent Neural Network*) [Sherstinsky, 2020] et RF

(*Random Forest*) [Biau et Scornet, 2016] qui ne peuvent fournir que la classification finale [Sui *et al*., 2021]. En effet, l'un des principaux points forts des modèles probabilistes est qu'ils donnent une idée de l'incertitude liée aux prédictions. Nous pouvons ainsi avoir une idée de la confiance d'un modèle de classification dans sa prédiction et l'interprétation des résultats sera plus efficace.

3.4.3 Démarche de l'approche ACSP

ACSP est une approche qui permet la prédiction de la réussite d'un contenu éducatif en ligne. Elle est d'autant plus générique et peut être adaptable pour résoudre le problème de prédiction de la réussite de différents contenus numérique avec des métadonnées et permettant de tracer les activités des utilisateurs. L'approche ACSP s'articule autour de 6 étapes comme le montre la Figure 3.6 ci-dessous.

- Étape 1 : Collecte des métadonnées des contenus éducatifs en ligne.
- Étape 2 : Préparation des données.
- Étape 3 : Construction du modèle de classification avec la régression logistique.
- Étape 4 : Evaluation de la performance du modèle.
- Étape 5 : Amélioration du modèle.
- Étape 6 : Prédiction de la réussite des contenus éducatifs.

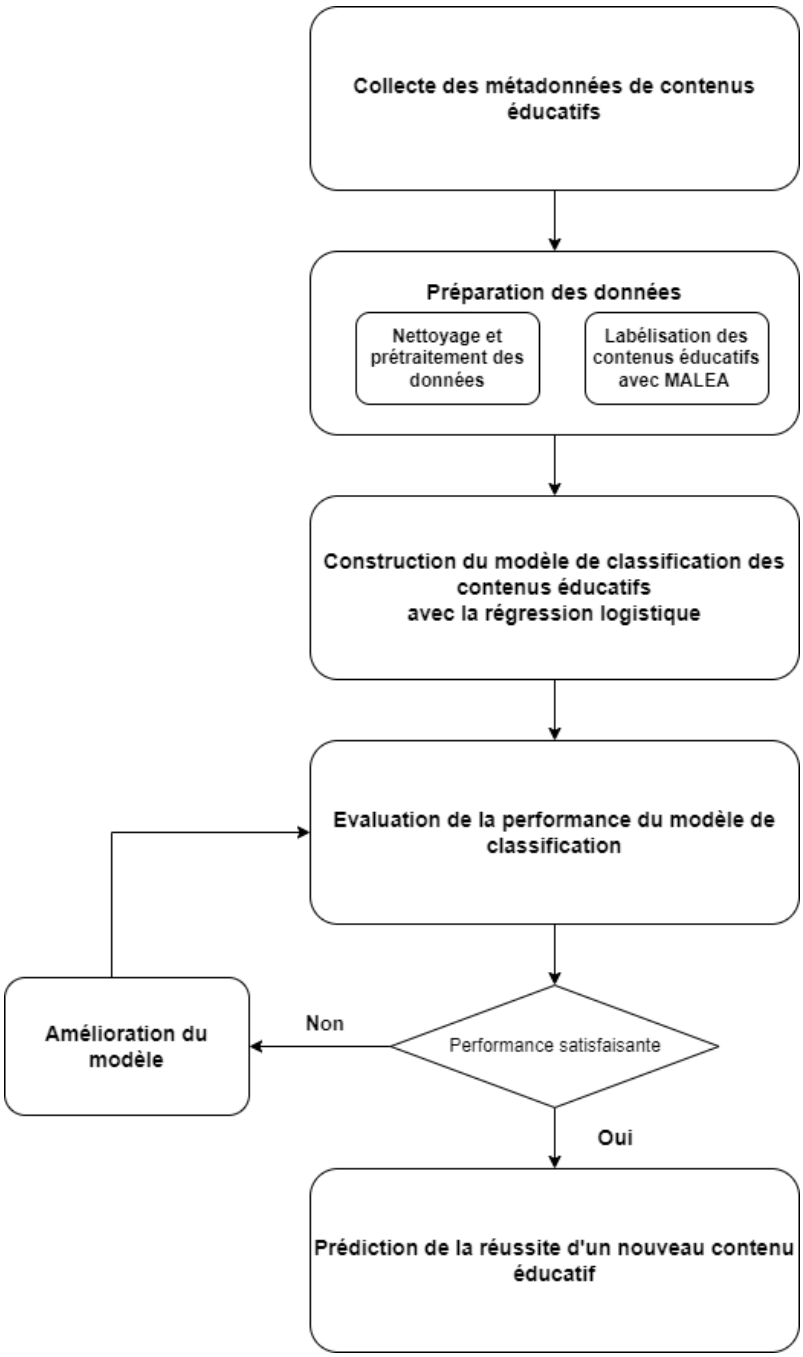

Figure 3.6 Processus de l'approche ACSP

Étape 1. Collecte des métadonnées des contenus éducatifs en ligne. Dans cette étape, nous collectons des données décrivant un ensemble de m contenus éducatifs en ligne sélectionnés à partir d'un EIAH.

Étape 2. Préparation des données. Dans cette étape, nous commençons par des opérations de nettoyage et de prétraitement des métadonnées des contenus éducatifs pour construire

l'ensemble $C(n, m)$. Cet ensemble est composé de n contenus éducatifs. Chaque contenu éducatif C_i *(i= 1…n)* est décrit et représenté par m caractéristiques $C_i = (c_1^i, c_2^i, \dots, c_m^i)$. Il sera, alors, question de sélection des m caractéristiques des contenus éducatifs, d'intégration, de transformation et de formatage de ces données. Une fois l'ensemble $C(n, m)$ préparé, nous passons à la phase de labélisation de cet ensemble pour créer la variable objectif *Y*. Ceci consiste à affecter à chaque contenu éducatif C_i un label y_i ayant pour valeur 1 ou 0 pour désigner respectivement un contenu éducatif réussi ou un contenu éducatif à améliorer. Pour assurer la tâche de labélisation, nous avons eu recours à l'approche MALEA. Le résultat de cette phase est présenté en Figure 3.7 ci-dessous.

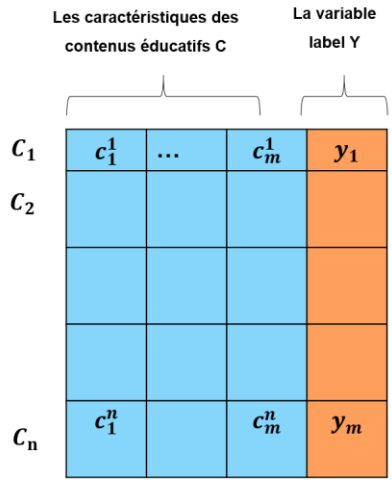

Figure 3.7 Le résultat final de l'étape préparation des données de l'approche ACSP

Étape 3. Construction du modèle de classification avec la régression logistique. Il s'agit de choisir l'algorithme adéquat capable de classifier les contenus éducatifs sous deux classes.
Une fois l'algorithme de régression logistique sélectionné, nous commençons l'apprentissage du modèle en utilisant des données d'entrainement/d'apprentissage. En effet, nous avons fractionné notre jeu de donnée en deux sous-ensembles : les données d'entrainement et les données de tests. Au cours de la phase d'entrainement, le modèle optimise ses paramètres, c'est-à-dire qu'il va trouver, par lui-même, les paramètres de la fonction *F* qui donne les meilleurs résultats. Pour optimiser ses paramètres, le modèle a besoin d'évaluer à chaque fois sa performance en mesurant, grâce à la fonction coût, l'erreur, celle-ci étant l'écart entre son résultat $\hat{Y}_i = F(C_i)$ et la vraie valeur du label Y_i *(i=1,…,k)*, avec k le nombre de contenus éducatifs de l'ensemble des données d'entrainements.

Étape 4. Evaluation de la performance du modèle. Au cours de cette étape nous évaluons notre modèle de régression logistique entrainé au cours de l'étape précédente. Pour ce faire, il faut le confronter à la réalité en utilisant les données de tests. Nous générons la matrice de confusion, pour pouvoir interpréter les résultats de la classification. Nous rappelons qu'avec la matrice de confusion les résultats sont classés sous quatre catégories : vrai positif, vrai négatif, faux positif et faux négatif. Cette classification nous permet de calculer la performance de notre modèle en utilisant plusieurs mesures comme la précision, le rappel, la spécificité et la sensitivité, mesures déjà évoquées à la section 2.4.2 du chapitre 2 et que nous mettrons en pratique dans le prochain chapitre.

Étape 5. Optimisation du modèle. Les prédictions/classifications avec la régression logistique sont déterminées d'après un seuil de probabilité d'appartenance d'un élément à l'une ou l'autre des deux classes générées par le modèle (dans notre cas, la classe des contenus éducatifs réussis ou la classe des contenus éducatifs à améliorer). Par défaut, le seuil standard de la classification par régression logistique est égal à 0,5 [Zhang *et al*., 2019]. Ainsi, la classification dépend du seuil. Dans cette étape nous avons effectué des variations au niveau du seuil, chacune est suivie par une évaluation de performance du modèle. Cette méthode d'ajustement du seuil nous a permis d'obtenir de meilleurs résultats.

Étape 6 : Prédiction de la réussite des contenus éducatifs. Une fois le modèle entrainé, puis amélioré, nous pouvons le déployer, afin qu'il puisse classer de nouveaux contenus éducatifs. La Figure 3.8 ci-après résume les trois phases principales de la prédiction de la réussite des contenus éducatifs en ligne, notamment l'apprentissage, le test et le déploiement.

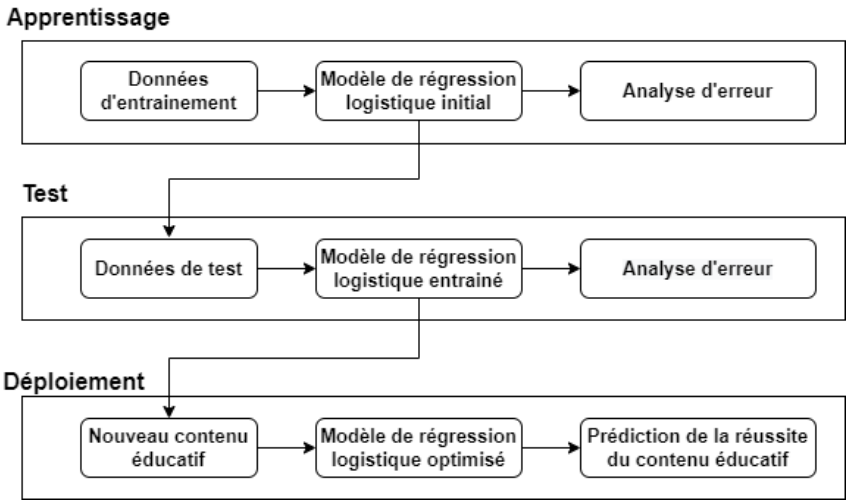

Figure 3.8 Architecture hiérarchique de la prédiction des contenus éducatifs en ligne

3.4.4 Classification binaire avec la régression logistique

La régression logistique est un algorithme de classification populaire en apprentissage automatique. Dans notre problème de classification, nous cherchons à attribuer une étiquette/label à une observation. Autrement dit, nous cherchons à attribuer une évaluation $\hat{y}_i$ à un contenu éducatif C_i.

La construction du modèle prédictif demande la provision d'un jeu de données comportant des contenus éducatifs labélisés. Ce jeu de données est construit par :

- une matrice $C(n,m)$ de *n* contenus éducatifs où chaque contenu est décrit par *m* caractéristiques

$$C(n,m) = \begin{pmatrix} c_1^1 & \dots & c_m^1 \\ & & \\ c_1^n & \dots & c_m^n \end{pmatrix}$$

- un vecteur $Y(n)$ de *n* labels

$$Y(n) = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix}$$

Le but est de trouver une ligne que nous appelons frontière de décision ou « *Boundary Decision* » en anglais, qui sépare les deux classes de contenus éducatifs, les contenus éducatifs réussis et les contenus éducatifs qui nécessitent des améliorations. Puisque la régression logistique est un modèle de classification linéaire, notre fonction hypothèse est la suivante :

$$S(C_i) = \beta_0 + \beta_1 c_1^i + \beta_2 c_2^i + \cdots + \beta_m c_m^i$$

Avec :

- C_i est l'une des n observations et qui correspond à un contenu éducatif ; cette variable est un vecteur qui contient $c_1^i, c_2^i, \dots c_m^i$
- c_j^i est l'une des m caractéristiques qui décrivent le contenu éducatif C_i et qui contribuent au calcul du modèle prédictif.
- β_j est un poids ou un paramètre de la fonction hypothèse. Nous cherchons à calculer ces paramètres pour obtenir notre fonction de prédiction.
- β_0 est une constante appelée biais.

Soit β le vecteur qui contient les paramètres $\beta_0, \beta_1, \ldots, \beta_m$. Notre fonction hypothèse est aussi le produit des deux vecteurs β et C.

$$S(C) = \beta C$$

Quand nous appliquons la fonction sigmoïde appelée aussi la fonction logistique σ sur notre fonction *S*, nous obtenons la fonction hypothèse pour la régression logistique donnée par :

$$H(C) = \sigma\big(S(C)\big) = \sigma(\beta C)$$

Avec la fonction σ, définie par :

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

Graphiquement, telle que représentée par la Figure 3.9 ci-dessous, la fonction σ correspond à une courbe en forme de S qui a pour limites 0 et 1 lorsque *x* tend respectivement vers -∞ et +∞.

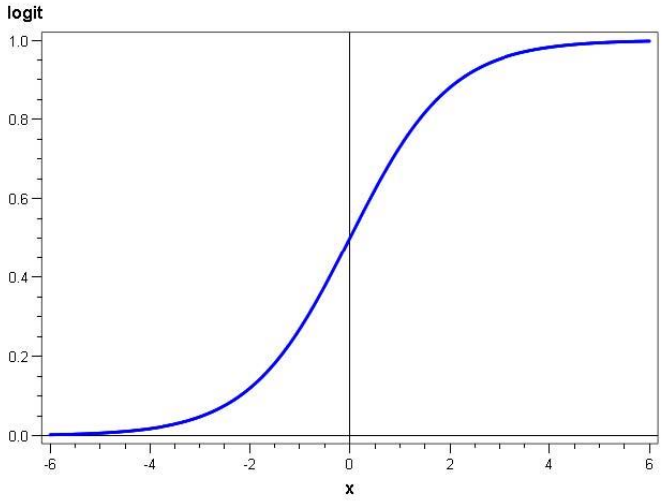

Figure 3.9 Courbe représentative de la fonction Sigmoïde

D'où la fonction qui définit la régression logistique, s'écrivant comme suit :

$$H(C) = \frac{1}{1+e^{-\beta C}}$$

Soient Cl_1 et Cl_2 les classes qui désignent respectivement les contenus éducatifs réussis et les contenus éducatifs à améliorer. Le modèle de régression logistique permet de prédire la probabilité qu'un contenu éducatif C_i soit réussi ($\hat{y}_i = 1$, c'est-à-dire $C_i \in Cl_1$) ou s'il nécessite des améliorations ($\hat{y}_i = 0$, c'est-à-dire $C_i \in Cl_2$) à partir de l'optimisation des paramètres de régression. Cette probabilité varie toujours entre 0 et 1. Lorsque la valeur prédite est inférieure à un seuil de classification, le contenu éducatif est susceptible d'avoir

besoin d'améliorations, alors que lorsque cette valeur est supérieure au même seuil, il n'en a pas besoin.

Si le modèle de régression logistique sait estimer ses paramètres $\beta_0, \beta_1, \ldots, \beta_m$ à partir des observations, la classification fournit la règle de décision suivante :

- C_i est classé contenu éducatif réussi si $S(C_i) > 0$, c'est-à-dire

$$p(Cl_1|C_i) = \sigma(S(C_i)) \geq \text{seuil de classification}$$

- C_i est classé contenu éducatif à améliorer si $S(C_i) < 0$, c'est-à-dire

$$p(Cl_2|C_i) = 1 - p(Cl_1|C_i) = \sigma(-S(C_i)) \geq \text{seuil de classification}$$

Ainsi, le problème de classification des contenus éducatifs par régression logistique est un problème d'optimisation où nous essayons d'obtenir le meilleur vecteur de paramètre $\beta = [\beta_0, \beta_1, \ldots, \beta_m]$ permettant à notre courbe sigmoïde d'associer au mieux les contenus éducatifs $C(n, m)$ aux classes $Y(n)$.

3.5 Conclusion

Au cours de ce chapitre, nous avons proposé un système d'aide à l'évaluation et l'amélioration des contenus éducatifs en ligne, basé sur deux approches appelées respectivement MALEA et ACSP.

L'approche MALEA (*Multicriteria Approach for Learning Experience Analysis*) est basée sur l'algorithme d'apprentissage non supervisé k-means. Cette approche est capable de regrouper les apprenants selon leur comportement. MALEA permet, plus précisément, d'identifier les groupes d'apprenants ayant des expériences d'apprentissage similaires. L'originalité de MALEA, par rapport aux méthodes existantes réside dans sa capacité d'utiliser plusieurs critères d'analyse. De plus, elle exploite des traces numériques des apprenants générés dans l'EIAH et non pas des données déclarées par les apprenants, sujettes à subjectivité. Ainsi, MALEA apporte de la précision au jugement humain.

L'approche ACSP (*Approach for Content Success Prediction*) combine l'approche MALEA et la méthode de régression logistique, afin de prédire la réussite des contenus éducatifs en ligne. ACSP offre deux atouts. Elle tient compte du comportement des apprenants dans le processus de l'évaluation des contenus éducatifs en ligne, d'une part. D'autre part, elle permet d'évaluer automatiquement les contenus éducatifs en ligne à tout moment, avant ou après leur diffusion sur l'EIAH.

Dans le chapitre qui suit, nous montrons, via deux cas d'étude réels, l'applicabilité de notre système d'aide à l'évaluation et l'amélioration des contenus éducatifs intégrant les deux approches MALEA et ACSP.

Chapitre 4 : Etudes expérimentales et analyse des résultats

4.1 Introduction

Dans ce chapitre, nous visons à confirmer nos contributions en effectuant deux types de validations : opérationnelle et expérimentale. D'une part, la validation opérationnelle vise à montrer que les deux approches proposées MALEA et ACSP agissent comme prévu et que le décideur pédagogique peut effectivement les utiliser (1) couplées dans un système d'aide à l'évaluation et l'amélioration des contenus éducatifs en ligne ou (2) séparément en fournissant à ACSP un jeu de données labélisées. D'autre part, la validation expérimentale permet de tester chacune de nos deux approches en utilisant des données expérimentales afin de montrer qu'elles fournissent les résultats attendus et vérifier leur performance à l'aide de différentes mesures.

À cet égard nous avons mené deux études de cas. Dans la première, nous mettons en œuvre le système d'aide à l'évaluation et l'amélioration des contenus éducatifs en ligne proposé dans le chapitre 3, avec des données réelles issues de la plateforme d'apprentissage en ligne Moodle de l'Université Virtuelle de Tunis (UVT). La mission principale de cet établissement d'enseignement public est de développer des cours et des programmes universitaires en ligne pour les universités publiques tunisiennes. Nous appliquons d'abord l'approche d'analyse multicritère des expériences d'apprentissage MALEA pour labéliser un ensemble de contenus éducatifs diffusés sur la plateforme Moodle de l'UVT. Ensuite, nous utilisons cet ensemble de contenus éducatifs labélisés pour construire, à l'aide de l'approche ACSP, un modèle prédictif capable de distinguer un contenu éducatif réussi d'un contenu éducatif nécessitant des améliorations que ce soit avant ou après sa diffusion sur l'EIAH.

La deuxième étude de cas consiste à appliquer l'approche MALEA avec des données collectées à partir du LMS Kalboard 360[15]. Cet ensemble de données, disponible gratuitement en ligne, est plus large (nombre d'observations et de caractéristiques) que celui de l'UVT, permettant d'explorer tous les critères d'analyse des expériences d'apprentissage identifiés dans ce travail de recherche.

[15] https://elearningindustry.com/directory/elearning-software/kalboard360

Nous suggérons, à la fin de ce chapitre, un ensemble de recommandations permettant aux concepteurs pédagogiques de l'UVT d'améliorer leurs contenus éducatifs en ligne. Tous les résultats obtenus sont analysés et discutés.

4.2 Première étude de cas basée sur les données générées par l'UVT

Cette section présente les deux études de cas menées dans le contexte de l'UVT pour valider opérationnellement le système d'aide à l'évaluation des contenus éducatifs en ligne proposé : (1) application de notre approche d'analyse multicritère des expériences d'apprentissage MALEA, (2) application de notre approche de prédiction de la réussite des contenus éducatifs en ligne ACSP.

Au regard de cette étude, nous nous proposons de suivre le scénario suivant :

Collecter un ensemble de 157 cours en ligne. Nous tenons à préciser que nous n'avons pas eu trop de choix en termes de la taille d'échantillonnage. Nous étions face à deux problèmes communs : la confidentialité des données et la restriction d'accès aux données.

- Pour chaque cours, extraire les traces des apprenants
- Pour chaque cours diffusé, appliquer l'approche MALEA dans le but :
    - d'analyser les expériences d'apprentissage des apprenants
    - d'attribuer un label représentant le résultat de son évaluation (réussi ou à améliorer)
- Créer une base de données des cours en ligne labélisés
- Appliquer ACSP pour prédire la réussite d'un nouveau cours en ligne avant sa diffusion.

4.2.1 Application de l'Approche d'analyse multicritère des expériences d'apprentissage (MALEA) : cas de l'UVT

Dans le cadre de ce travail de recherche, nous avons sélectionné un ensemble de 157 cours en ligne diffusés sur la plateforme Moodle[16] de l'UVT. Nous considérons qu'un cours en ligne est un contenu éducatif en ligne. Nous visons, donc, à attribuer à chacun de ces contenus éducatifs, un label représentant le résultat de son évaluation (*réussi* ou *à améliorer*). Comme nous l'avons

[16] https://iset.uvt.tn/

proposé, dans le chapitre 3, l'évaluation d'un contenu éducatif en ligne déjà diffusé sur l'EIAH est déterminée, en se basant sur l'analyse des expériences d'apprentissage des apprenants ayant suivi ce contenu, à l'aide de l'approche MALEA. Cette approche classe les apprenants dans des groupes par similarité. Le résultat du regroupement/ *clustering* convient pour créer les étiquettes dans les données qui sont ensuite utilisées pour l'entrainement/l'apprentissage de la machine [Hofmann, 2001]. Nous exposons, dans ce qui suit, un exemple d'application de l'approche MALEA dans lequel nous présentons les résultats des différentes étapes du processus d'évaluation (labélisation) d'un contenu éducatif en ligne. Pour ce faire, nous proposons le scénario suivant :

- Collecter les traces des apprenants ayant suivi le cours UML publié sur la plateforme de l'UVT.
- Préparer les données collectées.
- Regrouper les apprenants selon leurs expériences d'apprentissage selon deux groupes, à l'aide de l'algorithme k-means
- Identifier les caractéristiques communes des deux groupes d'apprenants : (1) ceux ayant des expériences d'apprentissage positives et (2) ceux ayant des expériences d'apprentissage négatives.
- Examiner la taille des deux clusters et attribuer un label/une évaluation du cours d'UML en se référant au taux des expériences d'apprentissage positives générées.
- Valider la performance de l'approche MALEA en comparant K-means à d'autres algorithmes de classification non supervisée.

4.2.1.1 Collecte des données

Comme annoncé auparavant, les données utilisées dans cette étude de cas sont collectées à partir de la plateforme d'apprentissage en ligne de l'Université Virtuelle de Tunis. L'UVT utilise la plateforme d'apprentissage en ligne Moodle. La confidentialité et le caractère privé des données personnelles sont respectés ; c'est pourquoi les données ne sont pas accessibles au public. Les données sont, cependant, disponibles auprès des administrateurs de la plateforme Moodle sur demande raisonnable et avec l'autorisation du directeur de l'UVT.

Dans cet exemple, nous recueillons des données relatives à des étudiants en deuxième année de licence informatique ayant suivi un cours d'UML en ligne. Les données obtenues ont différents formats, principalement des fichiers csv et des fichiers journaux tel qu'illustré dans les Figure 4.1, Figure 4.2, Figure 4.3.

C	D	E	F	G	H	I	J
Introduction du cours		Logiciel UML		Activité 1		Activité 2	
Not completed		Not completed		Not completed		Completed	Saturday,
Not completed		Not completed		Not completed		Completed	Saturday,
Not completed		Not completed		Not completed		Completed	Thursday,
Completed	Tuesday	Completed	Tuesday,	Completed	Tuesday	Completed	Saturday,
Not completed		Not completed		Not completed		Completed	Tuesday,
Completed	Saturday	Completed	Saturday	Completed	Saturday	Completed	Friday, 14
Not completed		Not completed		Not completed		Not completed	
Not completed		Not completed		Not completed		Completed	Tuesday,

Figure 4.1 Exemple de fichier csv contenant des données par rapport à l'achèvement des activités pédagogique

A	B	C	D	E	F	G
id	discussion	userid	created	modified	message	messageform
3	3	2	1492468974	1492468974	Que pensez-vous	1
4	3	31	1492511941	1492511941	j'aime beacoup cette	1
5	3	59	1492512301	1492512301	c'est une expÃ©r	1
6	3	21	1492514889	1492514889	i think its a good i	1
7	3	43	1492515583	1492515583	Je trouve que ce site	1
8	4	27	1492516155	1492516155	bien idÃ©e	1
9	3	44	1492517065	1492517065	je pense que c'est	1
10	4	9	1492520904	1492520904	tu veut dire bonn	1
11	3	9	1492521084	1492521084	j'ai aimÃ© cette e	1
12	3	22	1492521089	1492521089	c'est une bonne ir	1
13	3	18	1492548292	1492548292	belle idÃ©e..la cc	1

Figure 4.2 Exemple de fichier csv contenant des données par rapport à la participation des apprenants dans les forums de discussion

	A	B	C	D	E
	id	quiz	userid	grade	timemodified
	1	1	3	2.00000	1491389922
	2	1	31	8.00000	1492349856
	3	1	53	10.00000	1492949189
	4	1	9	9.00000	1492166141
	5	1	21	10.00000	1492177949
	6	1	40	10.00000	1492198981
	7	1	15	4.00000	1492520754
	8	1	6	10.00000	1492521378
	9	1	16	5.00000	1492248133
	10	1	13	5.00000	1492248812
	11	1	27	9.00000	1492248910
	12	1	58	10.00000	1492901511
	13	1	44	10.00000	1492249063

Figure 4.3 Exemple de fichier csv contenant les notes des apprenants dans les quiz

4.2.1.2 Préparation des données

Préalablement à la modélisation, les données ont été préparées, afin de les rendre compréhensibles par la machine. Ainsi, comme le montre la Figure 4.4, différents traitements ont été effectués dans cette étape, notamment la collecte, la sélection, le nettoyage, la transformation et le sauvegarde des données pour construire l'ensemble de données finales, qui seront transmises aux outils de l'analyse.

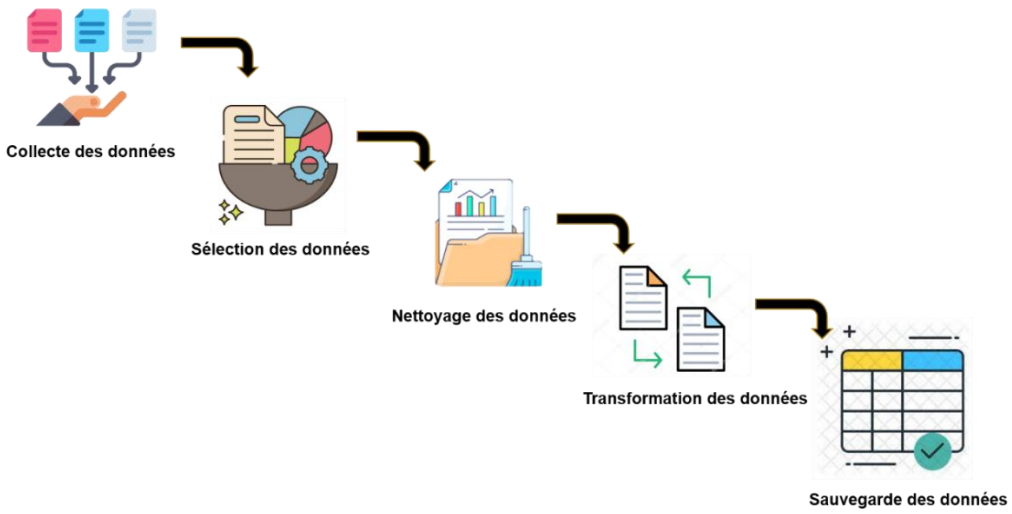

Figure 4.4 Préparation des données

L'identification des variables les plus représentatives constitue un élément décisif pour la réussite du regroupement des apprenants. À cet effet, les critères permettant d'analyser les expériences d'apprentissage, notamment l'engagement des apprenants, l'évaluation de la performance des apprenants et l'achèvement du contenu éducatif, ont d'abord été sélectionnés.

Les données permettant de mesurer chacun des trois critères sélectionnés ont ensuite été extraites. Ces données sont les variables qui vont être utilisées pour regrouper les apprenants selon leur expérience d'apprentissage. Le Tableau 4.1 ci-dessous décrit les critères et les variables que nous employons.

Tableau 4.1 Critères et variables d'analyse des expériences d'apprentissage

Critères	**Variables**	**Description**
Engagement des apprenants	n_course_access	Nombre d'accès au cours [Moubayed *et al*., 2020]
	n_forum_posts	Nombre de publications dans les forums de discussion [Ramesh *et al*., 2013],
	assignment_submission_ontime	Indicateur binaire sur la soumission des examens à temps
Evaluation de la performance des apprenants	Exam	La moyenne des notes dans les évaluations [Herodotou *et al*., 2019]
	f_result	Indicateur binaire sur le résultat final (réussite/ échec)
Achèvement du cours	assignment_submission	Soumission des tests et examens

Après l'identification des variables, il a été procédé à un nettoyage des données. Dans le cadre de celui-ci, une élimination de l'inconsistance a été opérée en traitant les valeurs manquantes et les données redondantes. Puis, les données ont été formatées en les transformant toutes sous format numérique comme illustré dans la Figure 4.5.

A	B	C	D	E	F	G
id_user	n_course_access	homework_submission	homework_submission_ontime	n_forum_posts	exam	f_result
11	179	1	0	0	4	0
12	204	1	0	0	4	0
13	22	1	1	0	4	0
14	193	1	1	0	5	1
15	141	1	0	0	5	1
16	64	1	1	0	5	1
17	15	1	1	0	5	1
18	104	1	1	0	6	1
19	39	1	1	0	6	1
20	62	1	1	0	6	1
21	57	1	1	1	6	1
22	3	1	1	1	7	1
23	129	1	1	1	7	1
24	85	1	1	1	7	1

Figure 4.5 Données nettoyées et transformées

4.2.1.3 Regroupement des apprenants avec l'algorithme k-means

Le regroupement des apprenants a été réalisé à l'aide de l'algorithme k-means. La première tâche à effectuer lors de la mise en œuvre d'un algorithme de regroupement (*clustering*) était de déterminer le nombre de clusters (k). Pour ce faire, nous avons eu recours à la méthode du coude (*Elbow*) [Madhulatha, 2012] dans laquelle l'algorithme k-means a été exécuté avec un nombre de clusters k variant de 1 à 10. L'idée de base consistait à définir k clusters en minimisant la variation intra-cluster (*within-cluster variation*). Par conséquent, nous avons calculé pour chaque valeur de k l'inertie ; celle-ci correspond à la somme des distances euclidiennes au carré entre chaque apprenant et le centroïde de son cluster. Pour déterminer le nombre optimal de clusters, il a fallu sélectionner la valeur de k au coude, c'est-à-dire le point après lequel l'inertie commence à diminuer de façon linéaire. Ainsi, selon la Figure 4.6, nous concluons que le nombre optimal de k est 2.

Nous avons fixé, dans la suite de cette étude de cas, la valeur à k=2. Ce choix est motivé non seulement par le fait qu'il représente la valeur optimale du nombre des clusters selon la méthode du coude, mais aussi par l'objectif métier de l'approche MALEA visant à évaluer le cours d'UML en ligne en diagnostiquant s'il nécessite des améliorations ou non.

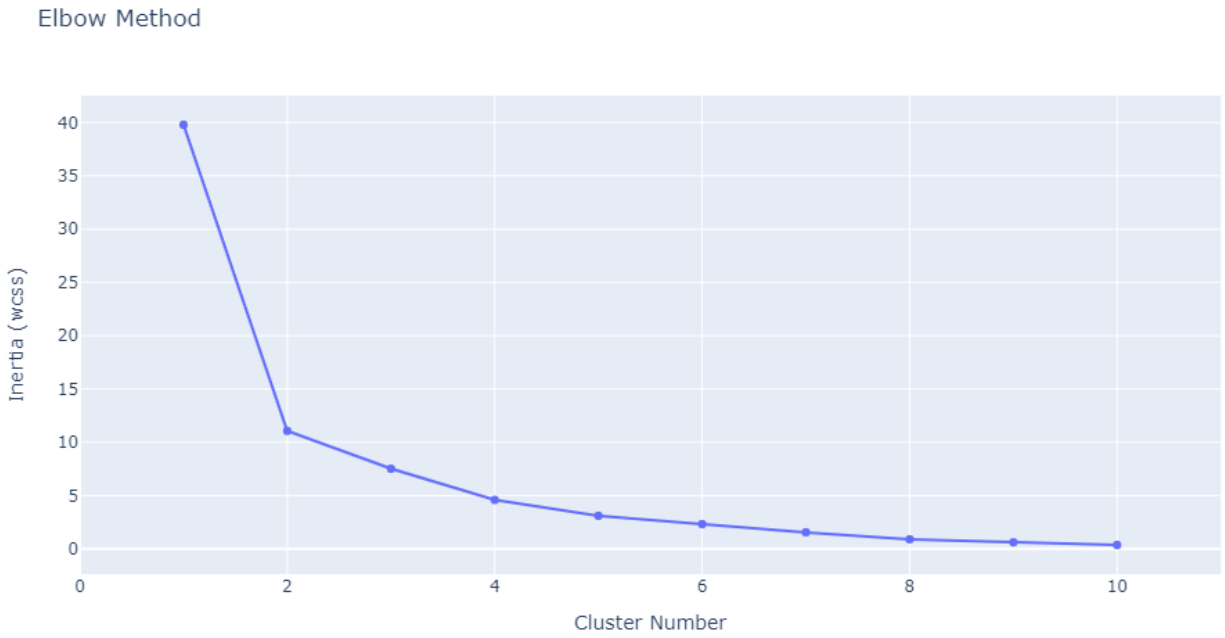

Figure 4.6 Identification du nombre de clusters optimal avec la méthode du coude

4.2.1.4 Identification des clusters

Dans notre cas d'étude, MALEA vise à analyser les expériences d'apprentissage des étudiants de la deuxième année de licence suivant le cours en ligne d'UML par l'observation de leurs comportements. L'application de l'algorithme k-means a donné lieu à un regroupement raisonnable. Comme le montre la Figure 4.7, deux clusters d'étudiants ont été identifiés. Le premier cluster, schématisé par la couleur bleue, représente les étudiants ayant des expériences d'apprentissage positives. Ces étudiants soumettent tous les travaux demandés et respectent généralement les délais. Ils participent parfois à des discussions ; pourtant, la culture des forums n'est pas très répandue dans la société tunisienne. Ces étudiants obtiennent de bonnes notes et réussissent l'examen final. Contrairement au premier, le deuxième cluster, schématisé par la couleur rouge, rassemble les étudiants ayant des expériences d'apprentissage négatives. Ces étudiants remettent rarement leurs travaux. Ils ne respectent généralement pas les délais, et ne participent jamais aux discussions. Ils sont très peu engagés par rapport aux étudiants du premier cluster et ont les plus mauvaises notes aux tests et examens.

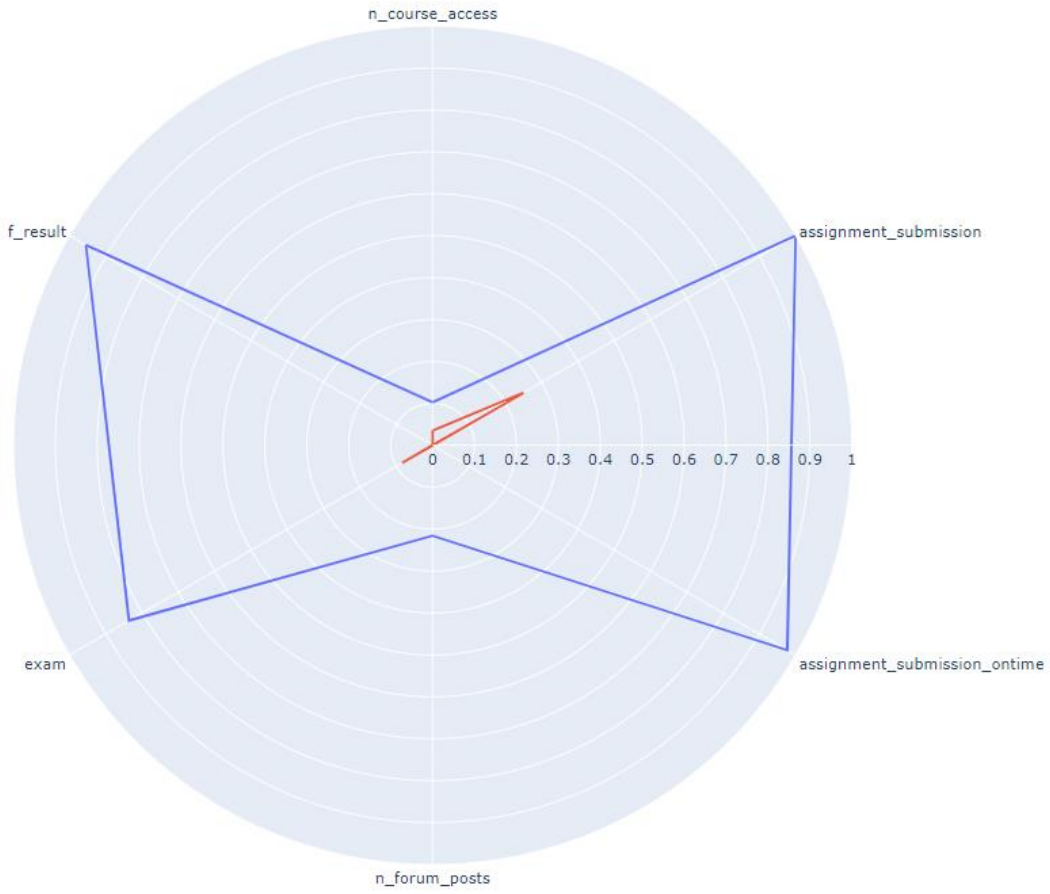

Figure 4.7 Résultat du groupement des apprenants avec le nombre de clusters k=2

4.2.1.5 Evaluation du contenu éducatif en ligne

Dans cette étape, nous proposons d'examiner la taille de chacun des deux clusters identifiés. Comme indiqué dans la Figure 4.8, les étudiants ayant des expériences d'apprentissage positives représentent 79% du nombre total des étudiants qui suivent le cours d'UML en ligne. Nous nous sommes basés sur ce pourcentage pour évaluer le cours en question à l'aide de l'échelle de mesure de Likert figurée dans le Tableau 4.2 [Gries *et al*., 2018] [Kangalgil et Özgül, 2018] [Linjawi et Alfadda, 2018] [Soykan et Kanbul, 2018] [Joshi *et al*., 2015].

Tableau 4.2 Evaluation des contenus éducatifs en ligne selon le taux des expériences d'apprentissage positives [Gries *et al*., 2018] [Kangalgil et Özgül, 2018] [Linjawi et Alfadda, 2018] [Soykan et Kanbul, 2018] [Joshi *et al*., 2015]

Evaluation de contenu éducatif en ligne	**Taux des expériences d'apprentissage positives**
Très faible	[0% - 20%[
Faible	[20% - 40%[
Moyen	[40% - 60%[
Fort	[60% - 80%[
Très fort	[80% - 100%]

Afin de décider si un contenu éducatif en ligne nécessite des améliorations, MALEA propose de fixer un seuil de réussite (α). En fait, la valeur de ce seuil peut être influencée par différents facteurs, principalement : le contexte de l'étude (intérêts des apprenants, spécificités culturelles, tendances de la politique pédagogique des prestataires du contenu éducatif, etc.) et les objectifs attendus. Dans notre contexte, nous avons fixé, après avoir consulté l'avis de différents experts pédagogiques impliqués dans le projet de l'UVT, le seuil de la réussite des cours en ligne diffusés par l'UVT à 60%. Par conséquent, nous pouvons considérer que le cours d'UML, sujet de cette étude de cas, est réussi.

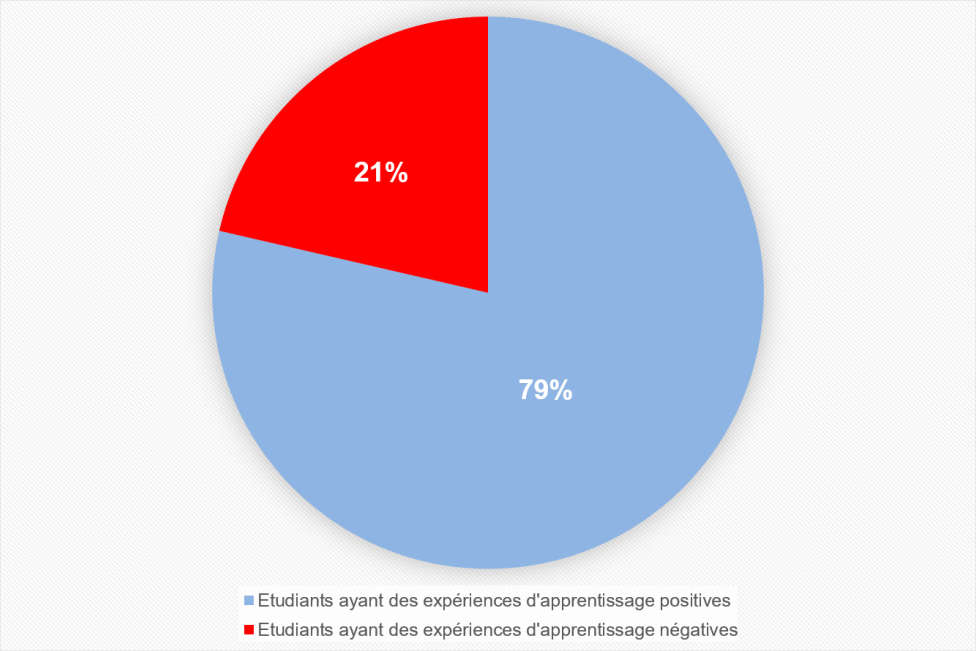

Figure 4.8 Distribution des étudiants dans les groupes identifiés

4.2.1.6 Validation de la performance de l'approche MALEA

Afin de valider notre approche MALEA expérimentalement, nous avons réalisé une analyse comparative basée sur l'étude de la qualité du groupement des apprenants. Pour ce faire, nous avons implémenté avec l'algorithme k-means, trois autres algorithmes de regroupement répandus issus de la littérature : la propagation d'affinité (*Affinity propagation*) [Laureano *et al.*, 2020] [Wang *et al.*, 2019] [Karga et Satratzemi, 2018], le partitionnement spectral (*Spectral clustering*) [Mengoni *et al*, 2018] [Cavallari *et al.*, 2017] et le regroupement aggloméré (*Agglomerative clustering*) [Hussain *et al.*, 2018] [Bharara *et al.*, 2018] [Hassel et Ridout, 2018]. Les plateforme Anaconda[17] et Google Colab[18] ont été utilisées comme environnements de développement. Les algorithmes ont été implémentés avec le langage Python à l'aide de la bibliothèque scikit-learn.

Nous avons évalué, ensuite, la performance du regroupement des quatre algorithmes en utilisant la méthode de la silhouette et l'indice de Davies Bouldin pour vérifier respectivement l'homogénéité des clusters et la séparation entre eux. Le coefficient de silhouette indique à quel point un étudiant est correctement affecté à son cluster. Ce coefficient appartient à l'intervalle [-1, 1]. Plus il est élevé et tend vers 1, plus le regroupement est satisfaisant. Le résultat de l'étude de la

[17] https://www.anaconda.com/products/distribution
[18] https://colab.research.google.com/

performance des quatre algorithmes est résumé dans le Tableau 4.3. Comme l'indique ce tableau, tous les algorithmes implémentés ont montré une performance satisfaisante. K-means a obtenu le coefficient de silhouette le plus élevé, égal à 0.609. Il est suivi par la propagation d'affinité avec un coefficient de silhouette égal à 0.607. À la troisième place, nous avons le regroupement aggloméré avec un coefficient de silhouette égal à 0.587. Le partitionnement spectral occupe la dernière place avec un coefficient de silhouette égal à 0.397. Ainsi, l'algorithme k-means maximise le coefficient de silhouette, ce qui montre la robustesse de notre approche proposée MALEA.
L'indice de Davies Bouldin indique que la meilleure séparation des clusters est obtenue avec les deux algorithmes k-means et regroupement aggloméré (0.049). S'ensuit la propagation d'affinité (0.354) puis le partitionnement spectral (0.745). Ainsi, le k-means et le regroupement aggloméré donnent également la meilleure performance. Nous avons retenu k-means car il est plus simple à mettre en œuvre, plus rapide et moins couteux en termes de mémoire de stockage que le regroupement aggloméré. Le regroupement aggloméré étant particulièrement utile lorsque l'objectif est d'organiser les clusters dans une hiérarchie, il ne répond pas à notre besoin (section 3.4.2).

Tableau 4.3 Evaluation de la performance des algorithmes k-means, propagation d'affinité, partitionnement spectral et regroupement aggloméré avec la méthode silhouette

Algorithme de regroupement	**Coefficient de silhouette moyen**	**Indice de Davies Bouldin**
k-means	0.609052	0.049862
propagation d'affinité	0.607449	0.354398
partitionnement spectral	0.397436	0.745086
regroupement aggloméré	0.587773	0.049862

4.2.2 Application de l'approche de prédiction de la réussite des contenus éducatifs en ligne ACSP

Dans cette section nous présentons les résultats de l'application de l'approche ACSP pour prédire la réussite des contenus éducatifs en ligne dans le contexte de l'UVT. Pour ce faire, nous proposons le scénario suivant :

- Collecter les 157 cours en ligne diffusés sur la plateforme de l'UVT.
- Préparer un jeu de données labélisé à partir des données collectées sur les cours. MALEA sera utilisée pour labélisation.
- Construire le modèle de classification permettant de distinguer un cours réussi d'un cours à améliorer en recourant à la régression logistique.
- Evaluer la performance du modèle à l'aide des mesures de performance : exactitude, rappel, spécificité et précision.
- Améliorer le modèle en ajustant le seuil de classification.
- Valider la performance de l'approche ACSP en comparant la régression logistique à d'autres algorithmes de classification binaire.

4.2.2.1 Collecte des métadonnées des contenus éducatifs en ligne

À travers la plateforme Moodle de l'UVT, nous avons sélectionné des cours en ligne à partir desquels nous avons créé le jeu de données de cette étude de cas. Comme expliqué dans le chapitre 3, ce jeu de données est composé de deux types de variables qui sont les caractéristiques décrivant les cours en ligne et le label permettant de distinguer si ce cours est réussi ou nécessite des améliorations. À cette étape, nous nous sommes chargés de collecter les caractéristiques des cours sélectionnés. Ces caractéristiques sont les suivantes [Mourali *et al*., 2021a] :

1. Description
2. Bande annonce
3. Durée (semaine)
4. Effort (h/semaine)
5. Planning
6. Objectif
7. Bénéficiaires du cours
8. Types des objets d'apprentissage
9. Nombre d'activités
10. Nombre de tests
11. Nombre de vidéos
12. Description des vidéos

13. Durée moyenne des vidéos
14. Nombre de vidéos de conférence
15. Nombre de photos
16. Nombre de quiz
17. Nombre de forum de discussion

4.2.2.2 Préparation des données

La première opération effectuée lors de cette phase a consisté à labéliser notre jeu de données. Cette opération a consisté à attribuer un label à chaque cours ayant la valeur 1, s'il s'agit d'un cours réussi, et la valeur 0, s'il s'agit d'un cours à améliorer. Ceci a été fait à l'aide de l'approche MALEA. Nous avons créé ainsi un jeu de données composé de 157 observations partagées entre 83 cours à améliorer et 74 cours réussis.

Une méthode pour surmonter le problème du manque de données d'entraînement est l'augmentation des données (*Data augmentation*) servant d'étendre l'ensemble de données en effectuant diverses manipulations [Ismael et Hefny, 2020]. Elle permet au modèle de mieux s'entraîner et améliorer la qualité de l'apprentissage, de réduire l'*overfitting* et d'augmenter les performances de généralisation [Ismael et Hefny, 2020]. L'augmentation des données résout le problème du déséquilibre des classes, en suréchantillonnant la classe minoritaire, ce qui permet d'obtenir un résultat équilibré sur le jeu de données d'apprentissage. Ainsi notre seconde opération a consisté à augmenter la quantité de nos données de façon à obtenir 233 observations. Pour ce faire, nous avons ajouté des copies légèrement modifiées de données. Cette technique a été utilisée pour enrichir l'ensemble de données d'entrainement d'un modèle de classification favorisant l'amélioration de sa performance.

Une fois les données finales prêtes, nous sommes passés à la dernière opération dans laquelle nous avons partitionné sous deux ensembles le jeu de données dont nous disposions. Nous avons réservé 70% du jeu de données (163 cours) au premier sous-ensemble devant être utilisé pour entrainer le modèle de classification ; le deuxième sous-ensemble, formé par 30% des données restantes (70 cours) devant servir pour tester le modèle de classification à des fins de prédiction de la réussite des contenus éducatifs en ligne.

4.2.2.3 Construction du modèle de classification avec la régression logistique

Les plateformes Anaconda et Google Colab ont été utilisées comme environnements de développement pour la construction du modèle de classification des cours en ligne de l'UVT. Les algorithmes de classification régression logistique, perceptron, adaline décente de gradient et adaline décente de gradient stochastique ont été implémentés avec le langage Python à l'aide de la bibliothèque *scikit-learn*. Nous avons évalué, ensuite, la performance des quatre algorithmes en utilisant la méthode de la validation croisée. Les résultats des évaluations sont résumés dans le Tableau 4.4 indiquant la robustesse du modèle de régression logistique.

Tableau 4.4 Evaluation de la performance du modèle de classification avec la méthode de la validation croisée

Algorithme de classification	**Validation croisée**
Régression logistique	0.84
Perceptron	0.66
Adaline décente de gradient	0.84
Adaline décente de gradient stochastique	0.74

4.2.2.4 Evaluation de la performance du modèle de prédiction

Nous avons construit, à cette étape, la matrice de confusion qui résume les résultats de prédiction de la réussite de nouveaux cours en ligne fournis par le sous-ensemble du test. La matrice de confusion illustrée dans le Tableau 4.5 compare les labels réels avec ceux prédits par notre modèle.

Tableau 4.5 Evaluation de la performance du modèle de prédiction avec la matrice de confusion

	Classe prédite : 0	**Classe prédite : 1**
Classe réelle : 0	Vrai Négatif (24)	Faux Positif (13)
Classe réelle : 1	Faux Négatif (3)	Vrai Positif (30)

Comme le montre le Tableau 4.6, nous pouvons, grâce à la matrice de confusion, calculer la performance du modèle de classification avec 4 métriques différentes, notamment l'exactitude, le rappel, la spécificité et la précision.

Tableau 4.6 Evaluation de la performance du modèle de classification avec la matrice de confusion

Métrique	**Définition**	**Formule**	**Performance**
Exactitude (*Accuracy*)	Elle permet de connaitre la proportion des prédictions correctes fournies par le modèle par rapport à toutes les prédictions	$\frac{(VP + VN)}{(VP + VN + FP + FN)}$	0,771
Rappel (*Recall*)	Il désigne la proportion des valeurs positives prédites avec précision	$\frac{VP}{(FN + VP)}$	0,909
Spécificité (*Specificity*)	Elle correspond au nombre de classes négatives prédites par le modèle.	$\frac{VN}{(VN + FP)}$	0,64
Précision (*Precision*)	Elle désigne la capacité d'un modèle à identifier uniquement les objets pertinents. Il s'agit du pourcentage de prédictions positives correctes.	$\frac{VP}{(VP + FP)}$	0,697

4.2.2.5 Amélioration du modèle de prédiction

L'objectif de l'application de l'approche ACSP consiste à aider les concepteurs pédagogiques à évaluer automatiquement leurs cours en ligne, afin de pouvoir les améliorer, si besoin. Augmenter la spécificité permet d'optimiser la détection des cours à améliorer. Minimiser les prédictions de type Faux Positifs (FP) est la solution que nous proposons, afin de remédier à ce problème. Pour ce faire, nous ajustons le seuil de la classification qui est, par défaut, égal à 0.5 [Zhang *et al*., 2019]. Nous présentons dans le Tableau 4.7 la matrice de confusion dans le cas où nous ajustons le seuil de la classification à 0.7 (après avoir testé un seuil égal à 0.8 et obtenu une spécificité maximale égale à 1 indiquant que le modèle est incapable de généraliser). Nous calculons la spécificité donnée par la formule suivante :

$$spécificité = \frac{VN}{(VN + FP)} = \frac{31}{(31 + 6)} = 0.83$$

Tableau 4.7 Evaluation de la performance du modèle de classification avec la matrice de confusion (seuil de classification=0.7)

	Classe prédite : 0	**Classe prédite : 1**
Classe réelle : 0	Vrai Négatif (31)	Faux Positif (6)
Classe réelle : 1	Faux Négatif (4)	Vrai Positif (29)

En ajustant le seuil de classification à 0.7, la spécificité s'est élevée de 0.64 à 0.84. Ceci est motivant pour fixer le seuil de classification de notre modèle prédictif à 0.7.

4.3 Seconde étude de cas : Application de l'approche MALEA dans le contexte de Kalboard 360

Nous avons mené cette deuxième étude de cas [Mourali *et al.*, 2021b], hors du contexte de l'UVT, afin de pouvoir appliquer notre approche MALEA sur un jeu de données plus large, non seulement en termes de nombre d'observations, mais aussi en termes de nombre de critères considérés dans l'analyse des expériences d'apprentissage. Nous détaillons ci-après le scénario suivi :

- Collecter les traces des apprenants ayant suivi le cours UML publié sur la plateforme de l'UVT.
- Préparer les données collectées.
- Regrouper les apprenants selon leurs expériences d'apprentissage sous deux groupes à l'aide l'algorithme k-means
- Identifier les caractéristiques communes des deux groupes d'apprenants : (1) ceux ayant des expériences d'apprentissage positives et (2) ceux ayant des expériences d'apprentissage négatives
- Identifier des modèles d'apprenants ayant besoin d'assistance spécifique et proposer des solutions pédagogiques pour les surmonter.

4.3.1 Collecte des données

Dans cette étude de cas, nous avons utilisé un ensemble de données éducatives disponible sur Kaggle[19]. Cet ensemble a été créé à partir de l'interaction des apprenants avec le système de gestion d'apprentissage (*Learning Management System*) Kalboard 360 [Aljarah, 2018]. Basé sur le cloud, Kalboard 360 est conçu pour aider des écoles à améliorer l'apprentissage/enseignement grâce à l'utilisation de nouvelles technologies d'éducation. Les données sont collectées à l'aide de l'outil expérience API (xAPI). Cet outil permet de suivre la progression de l'apprentissage, ainsi que des actions de l'apprenant, comme la lecture d'un article ou le visionnage d'une vidéo. Il aide les prestataires des formations en lignes à s'informer sur les expériences d'apprentissage des utilisateurs de Kalboard 360. L'ensemble de données contient des variables démographiques, académiques et comportementales de 480 apprenants.

4.3.2 Préparation des données

Nous avons commencé cette étape par la compréhension des données récoltées. Nous avons sélectionné d'après les critères et avons extrait les variables adéquates pour l'analyse des expériences d'apprentissage. Le Tableau 4.8 ci-dessous décrit les quatre critères considérés, ainsi que les variables employées dans notre analyse.

Tableau 4.8 Critères et variables d'analyse des expériences d'apprentissage

Critères	Variables	Description
Engagement	raisedhands	le nombre de fois où l'élève lève la main en classe
	VisITedResources	combien de fois l'élève visite un contenu éducatif
	AnnouncementsView	combien de fois l'élève consulte les nouvelles annonces
	Discussion	combien de fois l'élève participe à des groupes de discussion
Satisfaction	ParentAnsweringSurvey	les parents ont répondu ou non aux questionnaires fournis par l'école
	ParentschoolSatisfaction	le degré de satisfaction des parents
Achèvement	StudentAbsenceDays	le nombre de jours d'absence pour chaque élève (plus de 7, moins de 7)

[19] https://www.kaggle.com/c/student-academic-performance

Evaluation de la performance des apprenants	Assessment	la classification de l'élève par rapport à sa note totale (bas niveau, niveau moyen et haut niveau).

Ensuite, nous avons transformé les données de manière à obtenir un nouvel ensemble de données entièrement numériques. La Figure 4.9 montre un échantillon de l'ensemble de données transformées. Comme il n'y a pas de données manquantes, nous avons conservé les 480 observations du jeu de données initial.

	raisedhands	VisITedResources	AnnouncementsView	Discussion	ParentAnsweringSurvey	ParentschoolSatisfaction	StudentAbsenceDays	Assessment
0	15	16	2	20	1	1	0	2
1	20	20	3	25	1	1	0	2
2	10	7	0	30	0	0	7	1
3	30	25	5	35	0	0	7	1
4	40	50	12	50	0	0	7	2
5	42	30	13	70	1	0	7	2

Figure 4.9 Echantillon de l'ensemble de données transformées

4.3.3 Analyse des expériences d'apprentissage

Pour approfondir notre compréhension des données, une analyse préliminaire a été menée. Cette analyse a visé à étudier la dépendance entre les différentes variables sélectionnées. Pour ce faire, nous avons construit la matrice de corrélation visible en Figure 4.10. Celle-ci montre qu'il existe une forte corrélation entre le nombre de fois que l'apprenant visite un contenu de cours et sa note totale aux examens. Il existe, également, une corrélation entre le degré de satisfaction des parents et la réponse ou non aux enquêtes. Nous constatons que les parents qui répondent aux enquêtes sont dans la plupart des cas satisfaits. Contrairement, les parents insatisfaits ont tendance à ignorer les enquêtes.

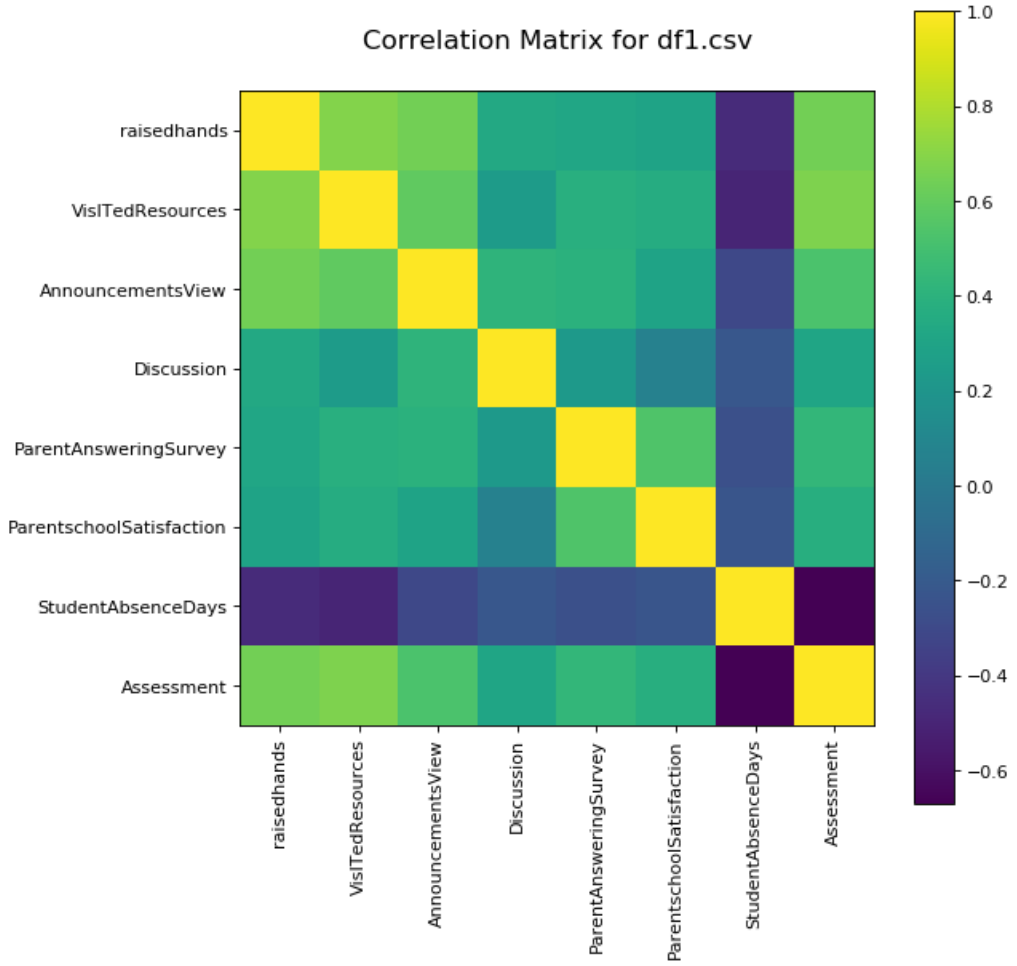

Figure 4.10 Etude de la dépendance entre les variables d'analyse des expériences d'apprentissage à l'aide de la matrice de corrélation

Après avoir examiné la corrélation entre les variables, nous avons appliqué l'algorithme k-means. L'objectif est de regrouper les apprenants selon l'expérience d'apprentissage, dans des clusters similaires. K-means a été exécuté avec un nombre de clusters *k* variant de 2 à 10. La qualité du regroupement a été ensuite évaluée par le coefficient de silhouette. La Figure 4.11 montre que l'augmentation du nombre de clusters a pour effet de dégrader la performance de l'algorithme. Nous avons fixé, alors, le nombre de k à 2, où le coefficient de silhouette est maximum (0.42).

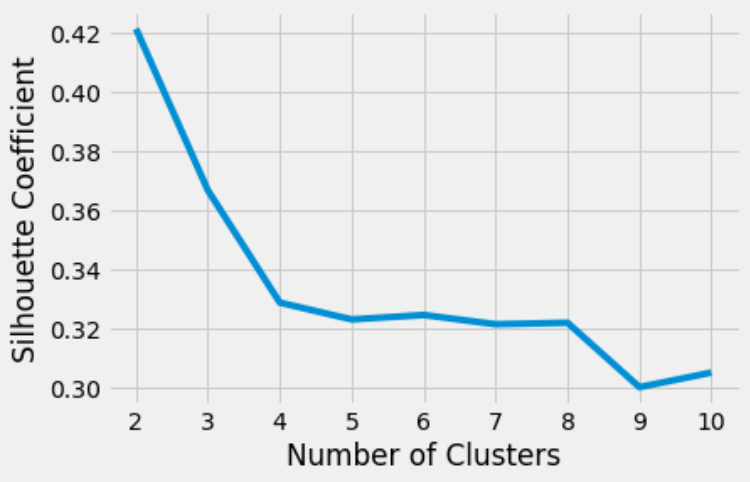

Figure 4.11 Coefficients de silhouette selon le nombre de clusters

4.3.4 Identification des clusters

L'application de l'algorithme k-means a permis d'identifier deux clusters d'apprenants, comme indiqué dans la Figure 4.12 ci-dessous. Le premier cluster, schématisé par la couleur rouge, représente les apprenants ayant des expériences d'apprentissage positives. Ces apprenants sont caractérisés par leur participation active à la fois en classe et dans les groupes de discussion. Ils visitent souvent le contenu éducatif et visualisent régulièrement les nouvelles annonces. Ils sont rarement absents et leurs parents sont généralement satisfaits. Quant au deuxième cluster, celui schématisé en bleu, il regroupe les apprenants ayant des expériences d'apprentissage négatives.

Figure 4.12 Distribution des étudiants dans les groupes identifiés

4.3.5 Résolution de problèmes éducatifs avec MALEA

À travers ce cas d'étude, nous montrons que MALEA est capable d'aider à comprendre comment les apprenants se comportent dans l'EIAH. Notamment, elle permet aux tuteurs, concepteurs pédagogiques et tout autre gestionnaire d'apprentissage en ligne utilisant des données éducatives, d'identifier des modèles d'apprenants ayant besoin d'assistance spécifique pour améliorer leurs expériences d'apprentissage. MALEA a permis dans cette étude de détecter les apprenants désengagés et ceux à risque d'obtenir de mauvais résultats. Elle a donné même des informations utiles sur les problèmes liés aux contenus éducatifs, impactant les expériences d'apprentissage.

Nous citons dans ce qui suit, des exemples de décisions qui pourraient être prises pour résoudre certains problèmes détectés grâce à l'application de MALEA.

D'abord, MALEA a montré que les apprenants ayant des expériences d'apprentissage négatives ne visitent pas, régulièrement, le contenu éducatif. C'est la raison pour laquelle ils obtiennent de mauvais résultats aux examens et aux tests d'évaluation. Pour résoudre ce problème, les concepteurs pédagogiques peuvent, par exemple, améliorer le contenu en intégrant des activités plus attrayantes. Des e-mails personnalisés peuvent être envoyés, aussi, aux apprenants afin de les motiver à accéder au contenu.

De plus, comme ces apprenants ne participent pas aux discussions, les tuteurs peuvent les engager dans des travaux collaboratifs avec un groupe de paires. Cela pourrait améliorer leur assiduité et leur intérêt [Awidi *et al*., 2019].

En outre, puisque les parents des apprenants ayant des expériences d'apprentissage négatives n'ont pas exprimé leur perception, une enquête peut leur être adressée. Ce serait l'occasion d'expliquer les raisons pour lesquels ils sont insatisfaits. Ces parents peuvent également être interrogés sur leurs attentes, propositions, etc.

4.4 Recommandations pour l'amélioration des cours en ligne

Nous tenons, dans cette partie, à proposer, aux concepteurs pédagogiques, un ensemble de recommandations pour les guider à améliorer la qualité de leurs cours en ligne, si nécessaire. Ces recommandations sont basées sur l'exploration des éléments caractérisant les cours en ligne réussis fournis par l'UVT.

Nous avons mené une étude statistique sur un échantillon de 74 cours en ligne classés réussis avec notre approche ACSP pour identifier les caractéristiques communes. D'après cette étude, un cours en ligne réussi doit tout d'abord être bien structuré tel montré dans la Figure 4.13. Il est crucial de commencer par une description du cours, de définir ses objectifs, présenter le planning et mentionner les informations à propos de la durée du cours et du nombre d'heures que l'apprenant doit consacrer par semaine.

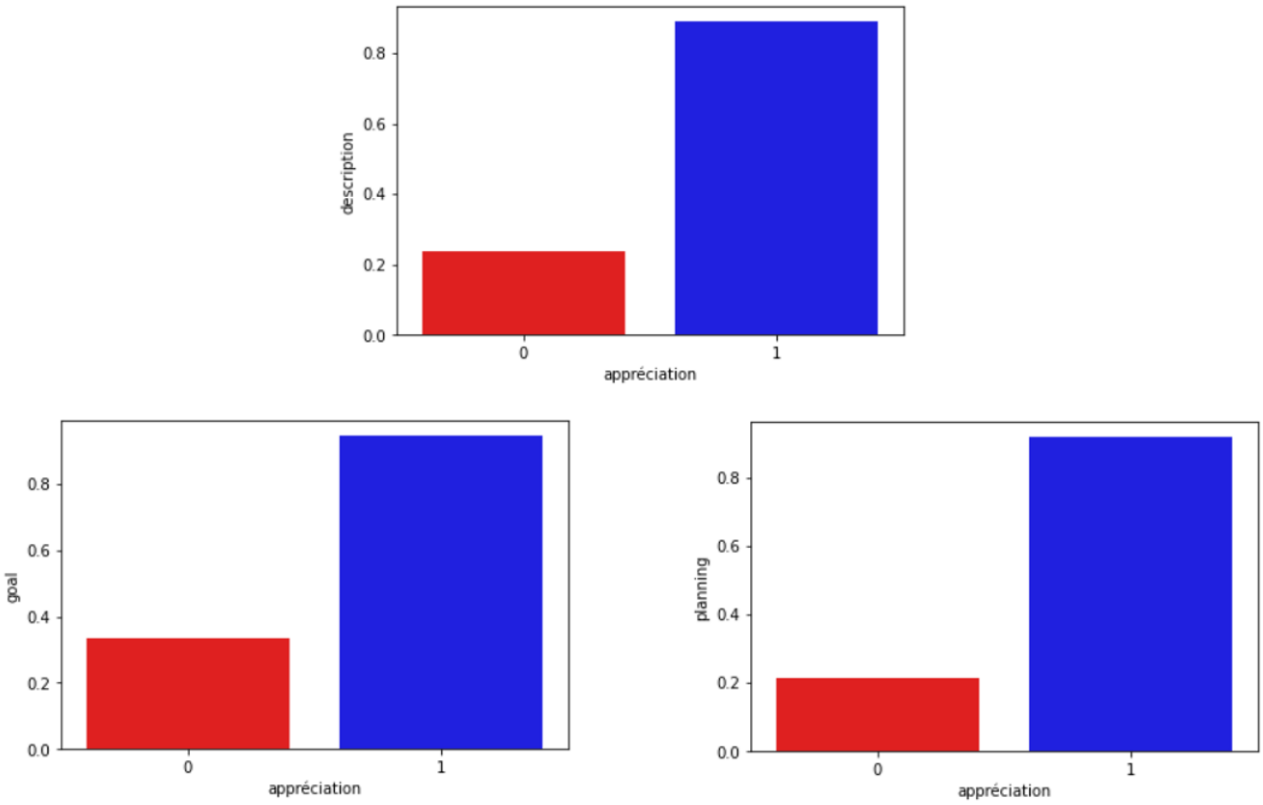

Figure 4.13 Importance de la description du cours, et présentation de ses objectifs et son planning

Un cours en ligne réussi doit, en outre, être diversifié et attractif. De plus, il est important d'inclure des contenus riches et interactifs. L'utilisation d'au moins 5 types d'objets d'apprentissage est souhaitée [Mourali *et al*., 2021a], comme illustré dans la Figure 4.14.

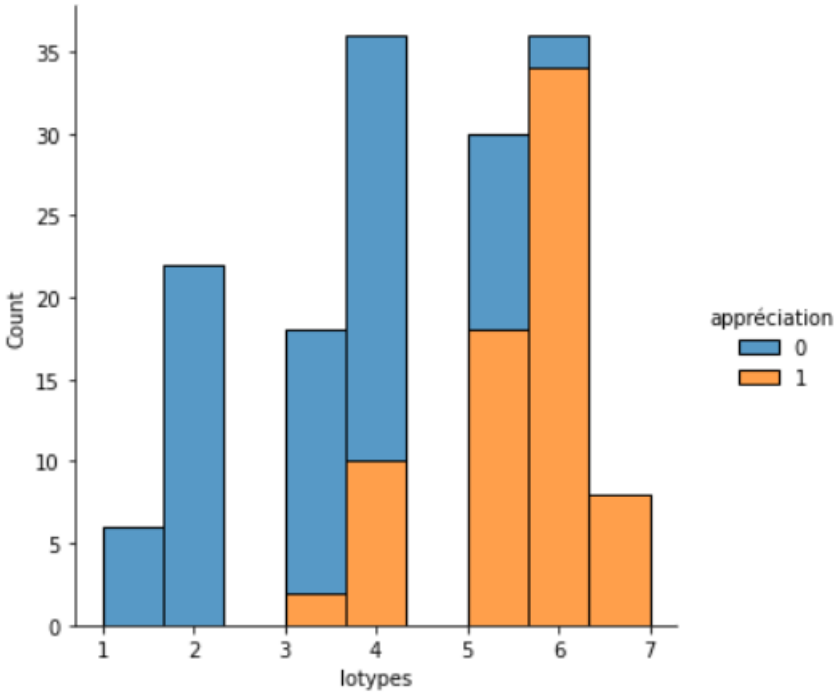

Figure 4.14 Types d'objet d'apprentissage utilisés dans les cours en ligne réussis de l'UVT

Selon la Figure 4.15, à partir de la répartition des types d'objets d'apprentissage utilisés dans les cours en ligne réussis de l'UVT, la vidéo occupe environ 30% du cours, alors que le texte et la photo occupent chacun environ 20-25% du cours. Pour éviter le risque d'anxiété face aux tests, le quiz est le meilleur moyen d'évaluer les apprenants. D'après cette étude, il s'est avéré que les cours réussis consacrent environ 10% du leur contenu aux quiz. Le reste du cours, qui représente près de 20%, peut prendre plusieurs autres formes tels que le liens, l'audio, l'animation, la simulation, etc. [Mourali *et al*., 2021a].

Un cours réussi doit être pratique et toujours présenté de manière pédagogique. À ce niveau, nous attirons l'attention sur la durée des vidéos, par exemple. La plupart des vidéos, partagées dans les cours en ligne réussis de l'UVT, ont une durée comprise entre 5 et 9 minutes.

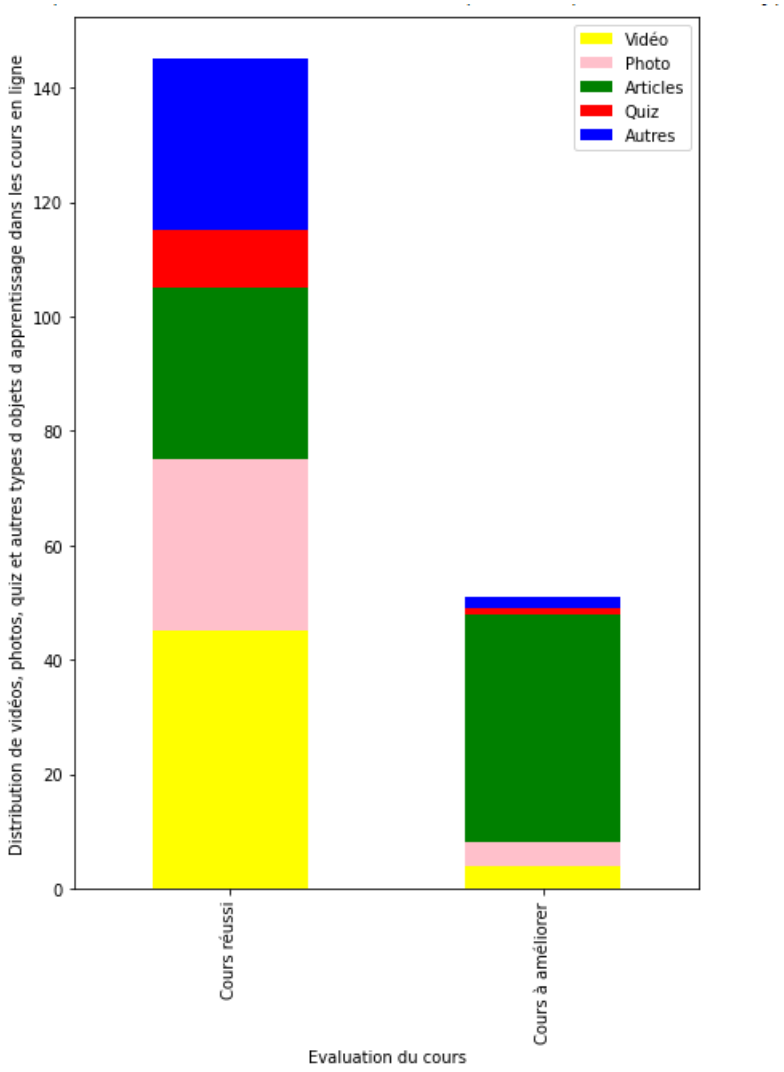

Figure 4.15 La répartition des types d'objet d'apprentissage utilisés dans les cours en ligne réussis de l'UVT

4.5 Conclusion

Au cours de ce chapitre, nous avons mené deux études de cas. Dans la première qui s'inscrit dans le cadre de l'UVT, nous avons d'abord appliqué notre approche MALEA pour analyser des expériences d'apprentissage selon trois critères, notamment l'engagement des apprenants, l'évaluation de la performance des apprenants et l'achèvement du cours en ligne. L'objectif de l'analyse était d'évaluer les cours en ligne de l'UVT.

Nous avons testé, en deuxième lieu, notre approche ACSP pour prédire la réussite des cours en ligne diffusés sur la plateforme Moodle de l'UVT.

Dans le deuxième cas d'étude, nous avons ajouté un quatrième critère pour vérifier la capacité de notre approche MALEA à supporter plus de critères. Pour ce faire, nous avons exploré les données éducatives du LMS Kalboard 360 qui nous a offert la possibilité d'inclure le critère de satisfaction. Les résultats obtenus dans le cadre de ces deux études présentées dans ce chapitre montrent que les approches MALEA et ACSP couplées répondent à l'objectif recherché et, ainsi peuvent être retenues pour l'évaluation et l'amélioration des contenus éducatifs en ligne. En outre, MALEA séparée est capable de fournir des informations utiles pour la prise de décisions dans le processus de résolution de problèmes éducatifs.

Plusieurs perspectives de recherche peuvent être envisagées suite à ces travaux. Nous les exposerons dans la conclusion générale de cette thèse.

Conclusion générale et perspectives

Conclusion générale

Dans un contexte d'apprentissage en ligne, la qualité du contenu éducatif est déterminante dans l'accessibilité au savoir et à la formation des compétences. Un contenu éducatif de qualité attire, retient, motive, fait progresser l'apprenant et justifie l'investissement des établissements éducatifs ou des entreprises dans le e-learning pour assurer une formation valable à un coût avantageux.

Nous ne pourrions affirmer qu'un contenu éducatif est valable qu'après l'avoir soumis à une évaluation. L'évaluation des contenus éducatifs en ligne a fait l'objet de plusieurs travaux de recherche qui, à notre connaissance, n'atteignent pas un degré d'objectivité et d'automatisation à notre avis satisfaisant.

Dans le cadre de cette thèse nous avons pu répondre aux quatre questions de recherche posées. Dans ce qui suit nous rappelons les contributions en rapport avec chaque question.

- (QR1) Comment évaluer objectivement les contenus éducatifs en ligne ?

  Pour répondre à cette question, nous avons effectué une évaluation automatique des contenus éducatifs en ligne à travers l'analyse des expériences d'apprentissage.

- (QR2) Quels critères peut-on retenir pour une analyse pertinente des expériences d'apprentissage ?

  En suivant une démarche d'investigation, nous avons pu identifier quatre critères utilisés dans l'état de l'art donnant la possibilité d'analyser et de comprendre les expériences d'apprentissage des apprenants : (1) engagement, (2) satisfaction, (3) performance des apprenants et (4) achèvement. Nous avons conclu cette investigation par une synthèse dans laquelle nous avons ressorti les variables permettant la mesure de chaque critère (section 1.3.2). Les deux hypothèses (H1) et (H2) annoncées dans l'introduction générale, ont été vérifiées en répondant à la troisième question et à la quatrième question.

- (QR3) Comment classifier les expériences d'apprentissage ?

  Pour répondre à cette question, nous avons proposé notre première contribution. Celle-ci consiste à développer une approche d'analyse multicritère des expériences d'apprentissage, dénommée MALEA (*Multicriteria Approach for Learning Experience Analysis*), sur

laquelle nous nous sommes appuyés pour évaluer des contenus éducatifs en ligne. L'apport de cette approche, par rapport aux autres méthodes disponibles dans la littérature, vise à surmonter l'imprécision du processus décisionnel humain afin d'améliorer l'objectivité des résultats. MALEA est basée sur l'algorithme d'apprentissage non supervisé k-means. Son objectif est d'identifier les groupes d'apprenants ayant des expériences d'apprentissage similaires. Pour atteindre cet objectif MALEA regroupe des apprenants en fonction de plusieurs critères que nous avons détaillés. MALEA repose sur cinq étapes : (1) la collecte des données des apprenants à travers leurs traces d'interaction, (2) la préparation de ces données, (3) le regroupement des apprenants dans des groupes similaires à l'aide de l'algorithme k-means, (4) l'identification des groupes d'apprenants et (5) l'évaluation du contenu éducatif en se basant sur le résultat du groupement.

- (QR4) Comment prédire le succès/la réussite d'un contenu éducatif en ligne ?

Pour répondre à cette question, nous avons proposé notre seconde contribution qui prend la forme d'une approche de prédiction de la réussite des contenus éducatifs en ligne. Elle se dénomme ACSP (*Approach for Content Success Prediction*). Cette dernière couple l'approche MALEA et la méthode de régression logistique issue de l'apprentissage automatique supervisé. Son objectif est de distinguer, automatiquement et avant qu'il ne soit diffusé, un contenu éducatif en ligne réussi, d'un autre qui nécessite des révisions. Pour atteindre cet objectif, nous avons appris à la machine à effectuer la tâche de classification binaire en lui fournissant un jeu de données labélisés. L'approche ACSP se décompose en six grandes étapes : (1) la collecte des métadonnées des contenus éducatifs en ligne, (2) la préparation du jeu de données y compris la labellisation à l'aide de MALEA, (3) la construction du modèle de classification binaire en se basant sur la régression logistique, (4) l'évaluation de la performance du modèle de classification, (5) l'amélioration du modèle de classification et (6) la prédiction de la réussite des contenus éducatifs.

Pour valider expérimentalement nos deux approches nous avons conçu un système d'aide à la décision pédagogique qui permet d'assister les concepteurs pédagogiques dans leurs tâches d'évaluation et d'amélioration de contenus éducatifs en ligne. Ce système combine MALEA et ACSP. Nous avons mené une étude de cas avec des données provenant de l'Université Virtuelle de Tunis (UVT), dont le but est d'évaluer des cours en ligne qu'ils soient diffusés ou prêts à être diffusés. À travers la plateforme Moodle de l'UVT, nous avons créé un jeu de données de 233

observations où chaque observation représente un cours diffusé sur la plateforme Moodle de l'UVT. Pour ce faire, nous avons collecté, dans un premier temps, les métadonnées des cours. Dans un deuxième temps, nous avons attribué à chaque cours un label grâce à l'approche MALEA. Ce label prend la valeur 1 s'il s'agit d'un cours réussi et la valeur 0 s'il s'agit d'un cours à améliorer. Pour déterminer le label d'un cours, nous avons appliqué MALEA qui permet d'analyser les expériences d'apprentissage des apprenants ayant suivi ce cours en se basant sur leurs traces d'interaction. Cette analyse est effectuée selon les trois critères : engagement de l'apprenant, achèvement du cours et évaluation de la performance de l'apprenant aux examens. Deux groupes d'apprenants ont été identifiés : un premier qui représente les apprenants ayant des expériences d'apprentissage positives et un deuxième qui représente les apprenants ayant des expériences d'apprentissage négatives. Selon la taille des deux groupes et en se référant à un seuil d'évaluation fixé après la consultation des experts pédagogiques, MALEA nous a permis de trouver le label d'évaluation de chacun des cours diffusés. Ainsi le jeu de données labélisé a été créé. Par la suite, nous avons construit, avec l'approche ACSP, le modèle de prédiction qui a été évalué et amélioré. Les résultats obtenus montrent que notre système d'aide à la décision pédagogique répond à l'objectif recherché et ainsi retenu pour l'évaluation des contenus éducatifs en ligne. D'un autre côté, afin de vérifier le résultat fourni, nous avons évalué, dans le cadre de ce système, la performance de MALEA à l'aide du coefficient de silhouette ainsi que la performance de l'approche ACSP en utilisant les quatre mesures de performance : exactitude, rappel, spécificité et précision.

Ainsi, l'approche MALEA et l'approche ACSP combinées permettent au concepteur pédagogique de procéder à l'évaluation automatisée du contenu éducatif à toute étape de son élaboration et notamment préalablement à sa diffusion sur l'Environnement Informatique pour l'Apprentissage Humain.

Perspectives

Dans cette thèse nous proposons des approches soutenant l'aide à la décision pédagogique. Plusieurs perspectives de recherche à la fois sur le plan théorique et le plan applicatif peuvent être envisagées, il pourrait s'agir de :

Perspectives à court terme

- Tester MALEA avec d'autres algorithmes de *clustering* comme k-means++ [Mydhili *et al.*, 2020] et k-médoïdes [Arora et Varshney, 2016] et comparer la performance des résultats. D'une part, k-means++ est une variante de k-means qui propose une initialisation intelligente des centroïdes. D'autre part, k-médoïdes est capable de rendre les centroïdes interprétables en les associant à des points de données réels.
- Etendre les expérimentations avec des jeux de données plus larges afin de tester l'efficacité de nos approches. D'une part, un jeu de données de traces des apprenants plus large en termes de nombre de caractéristiques serait souhaitable pour soutenir notre approche MALEA. La disponibilité de plus de caractéristiques comme le nombre d'interactions avec le tuteur [Reid, 2012], le nombre de ressources téléchargées par l'apprenant [Moubayed *et al.*, 2020], le temps écoulé dans le cours [Reid, 2012], etc., pourrait permettre d'approfondir l'analyse des expériences d'apprentissage. D'autre part, un jeu de données de contenus éducatifs plus large en termes de nombre d'observations donne l'occasion à la machine de mieux s'entraîner et améliorer ses performances dans la prédiction de la réussite des contenus éducatifs en ligne et ainsi de fournir de meilleurs résultats.
- Tester ACSP avec d'autres algorithmes de classification supervisée comme SVM, Naïve Bayes, arbre de décision, etc., et comparer la performance des résultats.

Perspectives à moyen terme

- Collaborer avec un concepteur pédagogique dans le but de créer un cours en ligne qui prend en compte les recommandations proposées et tester son efficacité.
- Faire une extension de l'approche ACSP en lui donnant la capacité de distinguer si un contenu éducatif a besoin de révisions mineures ou de révisions majeures. Pour ce faire nous visons à raffiner l'analyse des traces d'apprenants de minière et à donner un score au contenu plutôt qu'un label catégorique (réussi/à améliorer).
- Proposer une approche basée sur l'apprentissage profond (*Deep Learning*) permettant de recommander automatiquement des propositions aux concepteurs pédagogiques afin de les aider à améliorer leurs contenus éducatifs en ligne, telle que présentée ci-dessous dans la Figure 5.1. La première étape de cette approche consisterait à classifier les

contenus éducatifs sous deux catégories : « contenu réussi » et « contenu qui nécessite une révision » à l'aide de l'approche ACSP couplée avec MALEA. La deuxième étape serait consacrée à la recommandation automatique des améliorations du contenu éducatif. À cet effet, nous proposons le recours à l'apprentissage automatique profond pour apprendre à la machine à reconnaitre les caractéristiques d'un contenu éducatif en ligne réussi. Suivant ces caractéristiques de réussite, nous pourrions détecter les points faibles d'un nouveau contenu éducatif et recommander les améliorations adéquates.

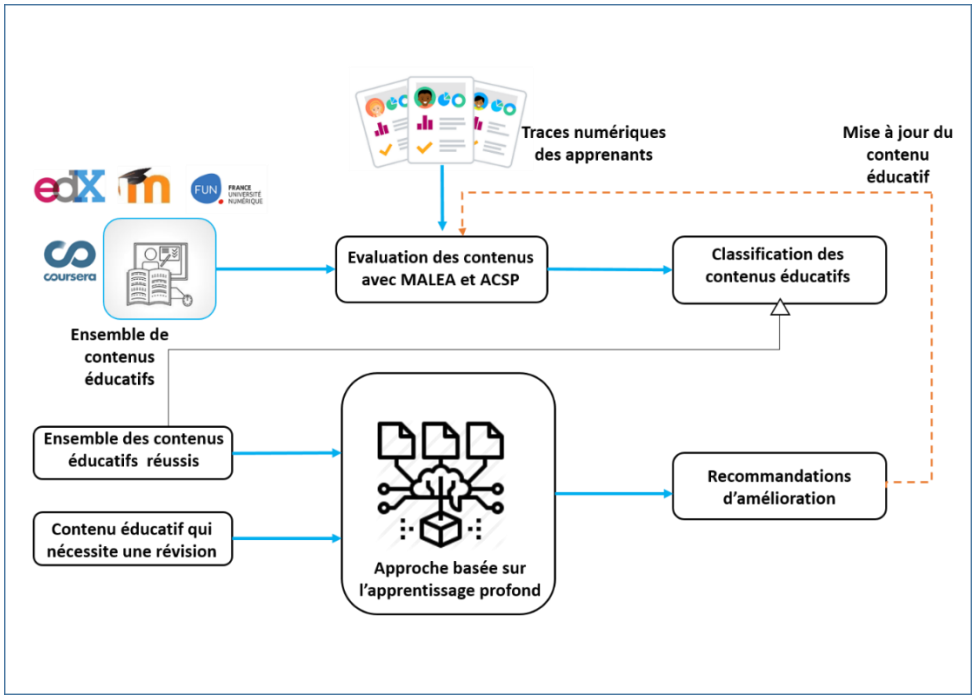

Figure 5.1 Approche basée sur l'apprentissage profond pour la recommandation d'amélioration des contenus éducatifs en ligne

- Formulation du problème :

  Etape 1 : Classification supervisée des contenus éducatifs en ligne avec l'approche ACSP (contenu réussi/ contenu qui nécessite une révision)

  Etape 2 : Recommandation automatique des améliorations à l'aide de l'apprentissage profond

  Tel que présenté dans la Figure 5.2, lors de cette étape, nous pourrions recourir à un réseau de neurones profond. Pour les données d'entrée nous disposerions d'un contenu éducatif à améliorer représenté par le vecteur $B\{b_1, \dots, b_k\}$ défini par ses

k caractéristiques b_i où ($i= 1 \ldots k$) et d'un ensemble de *m* contenus éducatifs réussis organisés dans une matrice *A* de dimension *(m,k)*. Chaque ligne de cette matrice représenterait un contenu éducatif $A_i \{a_1, \ldots, a_k\}$ défini par ses *k* caractéristiques/ attributs (durée, taille des vidéos, nombre de types des objets d'apprentissage utilisés, etc.). Notre objectif consisterait à comparer le profil d'un nouveau contenu éducatif à améliorer représenté par le vecteur $B\{b_1, \ldots, b_k\}$ avec le profil de chacun des contenus éducatifs réussis figurant dans la matrice *A*. Ceci signifie que les couches cachées du réseau de neurones profond devraient détecter au niveau du contenu éducatif $B\{b_1, \ldots, b_k\}$ les caractéristiques/ attributs ayant des valeurs insatisfaisantes par rapport aux autres contenus de la matrice *A*. Nous pourrions obtenir dans la couche de sortie, ainsi, une liste de recommandations d'amélioration à apporter aux attributs identifiés comme insatisfaisants.

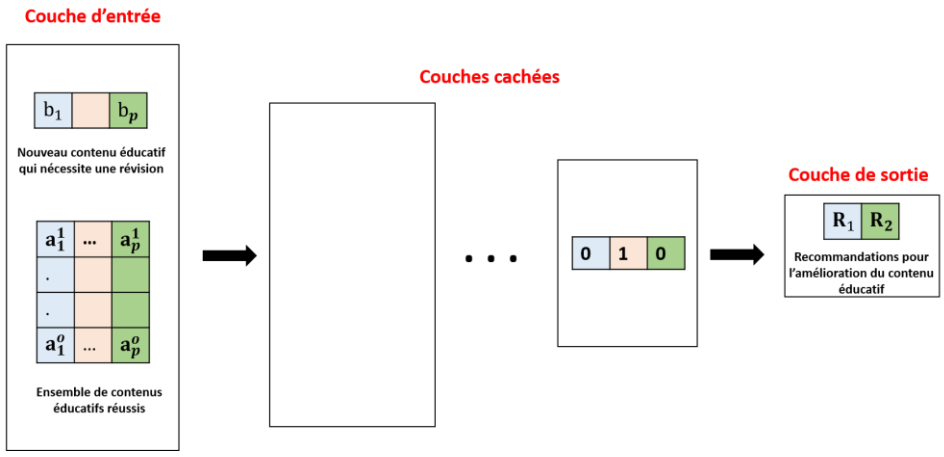

Figure 5.2 Approche basée sur l'apprentissage profond pour la recommandation des améliorations des contenus éducatifs en ligne

Perspectives à long terme

- Développer un système de personnalisation de l'apprentissage basé sur l'approche MALEA. Une façon possible d'aborder cette perspective est de ne plus faire l'évaluation du contenu par rapport à l'ensemble des apprenants, mais par groupe d'apprenants ayant certaines caractéristiques communes (par exemple un groupement par style d'apprentissage). Ainsi, il est possible de recommander des contenus sur cette base-là.

- Récupérer et réutiliser des évaluations déjà effectuées pour certaines parties des contenus.
- Tester l'intégralité de nos approches sur d'autres problématiques réelles d'entreprise afin de valider leur généralisation dans divers domaines comme l'e-commerce, les réseaux sociaux, la citoyenneté numérique, etc. À titre d'exemple, nous pourrions appliquer nos deux approches pour soutenir une campagne lancée sur les réseaux sociaux visant à sensibiliser les citoyens sur les questions environnementales. Comme il s'agit d'une action en ligne, des contenus explicatifs accrocheurs devraient être diffusés sur le web. Ces contenus devraient être capables de convaincre les citoyens et les engager pour adopter un comportement écoresponsable au quotidien. MALEA et ACSP couplées pourraient aider à diffuser des contenus pertinents qui prennent en considération les préférences, les orientations et les sensibilités des citoyens dans le but de réussir la campagne.

Références

[Abdelouarit *et al*., 2020] Abdelouarit, K. A., Sbihi, B., & Aknin, N. (2020, December). How Big Data Phenomenon Impact and Improve the e-Learning Process. In 2020 X International Conference on Virtual Campus (JICV) (pp. 1-4). IEEE.

[Aboagye *et al*., 2020] Aboagye, E., Yawson, J. A., & Appiah, K. N. (2020). COVID-19 and E-learning: The challenges of students in tertiary institutions. Social Education Research, 2(1), 1-8.

[Adam *et al*., 2018] Adam, K., Bakar, N. A. A., Fakhreldin, M. A. I., & Majid, M. A. (2018). Big data and learning analytics: a big potential to improve e-learning. Advanced Science Letters, 24(10), 7838-7843.

[Afify, 2018] Afify, M. K. (2018). E-learning content design standards based on interactive digital concepts maps in the light of meaningful and constructivist learning theory. JOTSE: Journal of Technology and Science Education, 8(1), 5-16.

[Agrebi *et al*., 2019] Agrebi, M., Sendi, M., & Abed, M. (2019, April). Deep reinforcement learning for personalized recommendation of distance learning. In World Conference on Information Systems and Technologies (pp. 597-606). Springer, Cham.

[Akter et Wamba, 2016] Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. Electronic Markets, 26(2), 173-194.

[Alamri *et al*., 2019] Alamri, A., Alshehri, M., Cristea, A., Pereira, F. D., Oliveira, E., Shi, L., & Stewart, C. (2019, June). Predicting MOOCs dropout using only two easily obtainable features from the first week's activities. In International Conference on Intelligent Tutoring Systems (pp. 163-173). Springer, Cham.

[Aldowah *et al*., 2020] Aldowah, H., Al-Samarraie, H., Alzahrani, A. I., & Alalwan, N. (2020). Factors affecting student dropout in MOOCs: a cause and effect decision-making model. Journal of Computing in Higher Education, 32(2), 429-454.

[Alharbi et al., 2016] Alharbi, Z., Cornford, J., Dolder, L., & De La Iglesia, B. (2016, July). Using data mining techniques to predict students at risk of poor performance. In 2016 SAI computing conference (SAI) (pp. 523-531). IEEE.

[Aljarah, 2018] Aljarah Ibrahim. Students' Academic Performance Dataset, xAPIEducational Mining Dataset, 2018.

[Alkhayrat et al., 2020] Alkhayrat, M., Aljnidi, M., & Aljoumaa, K. (2020). A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. Journal of Big Data, 7(1), 1-23.

[Almaiah et Alyoussef, 2019] Almaiah, M. A., & Alyoussef, I. Y. (2019). Analysis of the effect of course design, course content support, course assessment and instructor characteristics on the actual use of E-learning system. IEEE Access, 7, 171907-171922.

[Almohammadi et al., 2017] Almohammadi, K., Hagras, H., Alghazzawi, D., & Aldabbagh, G. (2017). A survey of artificial intelligence techniques employed for adaptive educational systems within e-learning platforms. Journal of Artificial Intelligence and Soft Computing Research, 7(1), 47-64.

[Alpaydin, 2020] Alpaydin, E. (2020). Introduction to machine learning. MIT press.

[Alshammari et al., 2018] Alshammari, S. H., Bilal Ali, M., & Rosli, M. S. (2018). LMS, CMS and LCMS: The confusion among them. Science International, 30(3), 455-459.

[Alzubi et al., 2018] Alzubi, J., Nayyar, A., & Kumar, A. (2018, November). Machine learning from theory to algorithms: an overview. In Journal of physics: conference series (Vol. 1142, No. 1, p. 012012). IOP Publishing.

[Anowar et al., 2021] Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). Computer Science Review, 40, 100378.

[Ansari et al., 2015] Ansari, Z., Azeem, M. F., Ahmed, W., & Babu, A. V. (2015). Quantitative evaluation of performance and validity indices for clustering the web navigational sessions. arXiv preprint arXiv:1507.03340.

[Arora et Varshney, 2016] Arora, P., & Varshney, S. (2016). Analysis of k-means and k-medoids algorithm for big data. Procedia Computer Science, 78, 507-512.

[Asoodar et al., 2016] Asoodar, M., Vaezi, S., & Izanloo, B. (2016). Framework to improve e-learner satisfaction and further strengthen e-learning implementation. Computers in Human Behavior, 63, 704-716.

[Avella et al., 2016] Avella, J. T., Kebritchi, M., Nunn, S. G., & Kanai, T. (2016). Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. Online Learning, 20(2), 13-29.

[Awidi et al., 2019] Awidi, I. T., Paynter, M., & Vujosevic, T. (2019). Facebook group in the learning design of a higher education course: An analysis of factors influencing positive learning experience for students. Computers & Education, 129, 106-121.

[Azcona et al., 2019] Azcona, D., Hsiao, I. H., & Smeaton, A. F. (2019). Detecting students-at-risk in computer programming classes with learning analytics from students' digital footprints. User Modeling and User-Adapted Interaction, 29(4), 759-788.

[Baghaee et al., 2019] Baghaee, H. R., Mlakić, D., Nikolovski, S., & Dragicević, T. (2019). Support vector machine-based islanding and grid fault detection in active distribution networks. IEEE Journal of Emerging and Selected Topics in Power Electronics, 8(3), 2385-2403.

[Balacheff, 2018] Balacheff, N. (2018). Les mots de la recherche sur les EIAH, enjeux et questions. Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation, 25(2), 63-94.

[Baruri et al., 2019] Baruri, R., Ghosh, A., Banerjee, R., Das, A., Mandal, A., & Halder, T. (2019, February). An empirical evaluation of k-Means clustering technique and comparison. In 2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 470-475). IEEE.

[Beckmann et al., 2015] Beckmann, M., Ebecken, N. F., & de Lima, B. S. P. (2015). A KNN undersampling approach for data balancing. Journal of Intelligent Learning Systems and Applications, 7(04), 104.

[Belarbi et al., 2018] Belarbi, N., Chafiq, N., Talbi, M., Namir, A., & Benlahmar, H. (2018, November). A recommender system for videos suggestion in a SPOC: A proposed personalized learning method. In International Conference on Big Data and Smart Digital Environment (pp. 92-101). Springer, Cham.

[Berry et al., 2019] Berry, M. W., Mohamed, A., & Yap, B. W. (Eds.). (2019). Supervised and unsupervised learning for data science. Unsupervised and Semi-Supervised Learning. Springer Nature. doi:10.1007/978-3-030-22475-.

[Bharara et al., 2018] Bharara, S., Sabitha, S., & Bansal, A. (2018). Application of learning analytics using clustering data Mining for Students' disposition analysis. Education and Information Technologies, 23(2), 957-984.

[Bhatore et al., 2020] Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. Journal of Banking and Financial Technology, 4(1), 111-138.

[Biau et Scornet, 2016] Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.

[Bourkoukou et El Bachari, 2018] Bourkoukou, O., & El Bachari, E. (2018). Toward a hybrid recommender system for e-learning personnalization based on data mining techniques. JOIV: International Journal on Informatics Visualization, 2(4), 271-278.

[Boutaba et al., 2018] Boutaba, R., Salahuddin, M. A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F., & Caicedo, O. M. (2018). A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. Journal of Internet Services and Applications, 9(1), 1-99.

[Broisin et Vidal, 2005] Broisin, J., & Vidal, P. (2005). Un environnement informatique pour l'apprentissage humain au service de la virtualisation des objets pédagogiques. Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation, 12(1), 177-204.

[Bussmann et al., 2021] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. Computational Economics, 57(1), 203-216.

[Cai et al., 2018] Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70-79.

[Camargo et al., 2020] Camargo, C. P., Tempski, P. Z., Busnardo, F. F., Martins, M. D. A., & Gemperli, R. (2020). Online learning and COVID-19: a meta-synthesis analysis. Clinics (Sao Paulo, Brazil), 75, e2286.

[Catal et al., 2019] Catal, C., Kaan, E. C. E., Arslan, B., & Akbulut, A. (2019). Benchmarking of regression algorithms and time series analysis techniques for sales forecasting. Balkan Journal of Electrical and Computer Engineering, 7(1), 20-26.

[Cavallari et al., 2017] Cavallari, S., Zheng, V. W., Cai, H., Chang, K. C. C., & Cambria, E. (2017, November). Learning community embedding with community detection and node embedding on graphs. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 377-386).

[Ceron et Iacus, 2016] Ceron, A., Curini, L., & Iacus, S. M. (2016). Politics and big data: Nowcasting and forecasting elections with social media. Routledge.

[Charbuty et Abdulazeez, 2021] Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20-28.

[Chauhan et al., 2019] Chauhan, V. K., Dahiya, K., & Sharma, A. (2019). Problem formulations and solvers in linear SVM: a review. Artificial Intelligence Review, 52(2), 803-855.

[Chen, 2018] Chen, H. (2018). Personalized recommendation system of e-commerce based on big data analysis. Journal of Interdisciplinary Mathematics, 21(5), 1243-1247.

[Chollet, 2017] Chollet, F. (2017). Deep Learning with Python. USA: Manning Publications Co, pp.5. https://dl.acm.org/doi/book/10.5555/3203489

[Chui et al., 2020] Chui, K. T., Fung, D. C. L., Lytras, M. D., & Lam, T. M. (2020). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. Computers in Human Behavior, 107, 105584.

[Cisel, 2016] Cisel, M. (2016). Utilisations des MOOC: éléments de typologie (Doctoral dissertation, Université Paris-Saclay).

[Clarke, 2013] Clarke, T. (2013). The advance of the MOOCs (massive open online courses): The impending globalisation of business education?. Education+ Training.

[Class Central, 2022] Class Central, (2022). MOOCs by the numbers in 2021. Available from https://www.classcentral.com/report/mooc-stats-2021/ [Online]. [Accessed: 10-Jul-2022]

[Clavel, 2019] Clavel, C. (2019). Analyse des opinions dans les interactions : De la fouille de données à l'interaction humain-agent. ISTE Group.

[Costa et al., 2019] Costa, L., Souza, M., Salvador, L., & Amorim, R. (2019, July). Monitoring students performance in e-learning based on learning analytics and learning educational objectives. In 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT) (Vol. 2161, pp. 192-193). IEEE.

[Cunningham, 2017] Cunningham, I. (2017). The wisdom of strategic learning: The self managed learning solution. Routledge.

[Dalipi et al., 2018] Dalipi, F., Imran, A. S., & Kastrati, Z. (2018, April). MOOC dropout prediction using machine learning techniques: Review and research challenges. In 2018 IEEE Global Engineering Education Conference (EDUCON) (pp. 1007-1014). IEEE.

[Daradoumis et al., 2013] Daradoumis, T., Bassi, R., Xhafa, F., & Caballé, S. (2013, October). A review on massive e-learning (MOOC) design, delivery and assessment. In 2013 eighth international conference on P2P, parallel, grid, cloud and internet computing (pp. 208-213). IEEE.

[Dash et al., 2019] Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. Journal of Big Data, 6(1), 1-25.

[Davies et Bouldin, 1979] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, (2), 224-227.

[Deepa et al., 2021] Deepa, N., Prabadevi, B., Maddikunta, P. K., Gadekallu, T. R., Baker, T., Khan, M. A., & Tariq, U. (2021). An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier. The Journal of Supercomputing, 77(2), 1998-2017.

[De Freitas et al., 2015] De Freitas, S. I., Morgan, J., & Gibson, D. (2015). Will MOOCs transform learning and teaching in higher education? Engagement and course retention in online learning provision. British journal of educational technology, 46(3), 455-471.

[Dhahri et al., 2019] Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W., & Faisal Nagi, M. (2019). Automated breast cancer diagnosis based on machine learning algorithms. Journal of healthcare engineering, 2019, 4253641. https://doi.org/10.1155/2019/4253641.

[Dieumegard et Durand, 2005] Dieumegard, G., & Durand, M. (2005). L'expérience des apprenants en e-formation: revue de littérature. Savoirs, (1), 93-109.

[Dimitrov, 2016] Dimitrov, D. V. (2016). Medical internet of things and big data in healthcare. Healthcare informatics research, 22(3), 156-163.

[Dogbe-Semanou, 2016] Dogbe-Semanou, D. A. K. (2016). Persévérance et abandon des apprenants à distance en Afrique subsaharienne francophone: cas du Togo (Doctoral dissertation, Université de Lomé (Togo)).

[ElSayed et al., 2019] ElSayed, A. A., Caeiro-Rodríguez, M., MikicFonte, F. A., & Llamas-Nistal, M. (2019, September). Research in learning analytics and educational data mining to measure self-regulated learning: A systematic review. In World conference on mobile and contextual learning (pp. 46-53).

[Farhat et al., 2020] Farhat, R., Mourali, Y., Jemni, M., & Ezzedine, H. (2020, February). An overview of Machine Learning Technologies and their use in E-learning. In 2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA) (pp. 1-4). IEEE.

[Fatima and Pasha, 2017] Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications, 9(01), 1.

[Felder et Silverman, 1988] Felder, R. M., & Silverman, L. K. (1988). Learning and teaching styles in engineering education. Engineering education, 78(7), 674-681.

[Feng et al., 2019] Feng, W., Tang, J., & Liu, T. X. (2019, July). Understanding dropouts in MOOCs. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 517-524).

[Foster et Siddle, 2020] Foster, E., & Siddle, R. (2020). The effectiveness of learning analytics for identifying at-risk students in higher education. Assessment & Evaluation in Higher Education, 45(6), 842-854.

[François-Lavet et al., 2018] François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., & Pineau, J. (2018). An introduction to deep reinforcement learning. Foundations and Trends® in Machine Learning, 11(3-4), 219-354.

[Frey et Dueck, 2007] Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. Science, 315(5814), 972-976.

[Garcia et al., 2018] Garcia, R., Falkner, K., & Vivian, R. (2018). Systematic literature review: Self-Regulated Learning strategies using e-learning tools for Computer Science. Computers & Education, 123, 150-163.

[Gaussier et Yvon, 2011] Gaussier, E., Yvon, F. (2011). Modèles statistiques pour l'accès à l'information textuelle. Hermès / Lavoisier, pp.482, 2011, ISBN 10 : 2746224976; ISBN 13 : 9782746224971.

[Gedrimiene et al., 2020] Gedrimiene, E., Silvola, A., Pursiainen, J., Rusanen, J., & Muukkonen, H. (2020). Learning analytics in education: Literature review and case examples from vocational education. Scandinavian Journal of Educational Research, 64(7), 1105-1119.

[Geisslinger, 2019] Geisslinger, M. (2019). Autonomous driving:“object detection using neural networks for radar and camera sensor fusion (Doctoral dissertation, Technical University of Munich, Allemagne).

[George, 2001] George, S. (2001). Apprentissage collectif à distance, SPLACH: un environnement informatique support d'une pédagogie de projet (Doctoral dissertation, Université du Maine, France).

[Georgescu et Bogoslov, 2019] Georgescu, M. R., & Bogoslov, I. A. (2019, November). Importance and Opportunities of Sentiment Analysis in Developing E-Learning Systems through Social Media. In DIEM: Dubrovnik International Economic Meeting (Vol. 4, No. 1, pp. 83-93). Sveučilište u Dubrovniku.

[Ghribi et al., 2010] Ghribi, M., Cuxac, P., Lamirel, J. C., & Lelu, A. (2010, January). Mesures de qualité de clustering de documents: Prise en compte de la distribution des mots clés. In 10ième Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances-EGC 2010.

[González, 2017] González, R. J. (2017). Hacking the citizenry?: Personality profiling, ‘big data’ and the election of Donald Trump. Anthropology Today, 33(3), 9-12.

[Gonçalves et al., 2018] Gonçalves, A. L., Carlos, L. M., da Silva, J. B., & Alves, G. R. (2018, June). Personalized Student Assessment based on Learning Analytics and Recommender Systems. In 2018 3rd International Conference of the Portuguese Society for Engineering Education (CISPEE) (pp. 1-7). IEEE.

[Goodfellow et al., 2017] Goodfellow, I., Bengio, Y., & Courville, A. (2017). Deep learning (adaptive computation and machine learning series). Cambridge Massachusetts, 321-359.

[Gregori et al., 2018] Gregori, P., Martínez, V., & Moyano-Fernández, J. J. (2018). Basic actions to reduce dropout rates in distance learning. Evaluation and program planning, 66, 48-52.

[Gries et al., 2018] Gries, K., Berry, P., Harrington, M., Crescioni, M., Patel, M., Rudell, K., Safikhani, S., Pease, S., & Vernon, M. (2018). Literature review to assemble the evidence for response scales used in patient-reported outcome measures. Journal of patient-reported outcomes, 2(1), 1-14.

[GuangJun et al., 2020] GuangJun, L., Nazir, S., Khan, H. U., & Haq, A. U. (2020). Spam detection approach for secure mobile message communication using machine learning algorithms. Security and Communication Networks, 2020,vol. 2020, ArticleID 8873639, 6 pages, 2020.
https://doi.org/10.1155/2020/8873639

[Haghighi et al., 2018] Haghighi, S., Jasemi, M., Hessabi, S., & Zolanvari, A. (2018). PyCM: Multiclass confusion matrix library in Python. Journal of Open Source Software, 3(25), 729.

[Hair Jr et Sarstedt., 2021] Hair Jr, J. F., & Sarstedt, M. (2021). Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing. Journal of Marketing Theory and Practice, 29(1), 65-77.

[Hajjar, 2014] Hajjar, C. (2014). Cartes auto-organisatrices pour la classification de données symboliques mixtes, de données de type intervalle et de données discrétisées (Doctoral dissertation, Supélec, France).

[Halevy et al., 2009] Halevy, Alon, Peter Norvig, and Fernando Pereira. "The unreasonable effectiveness of data." IEEE intelligent systems 24.2 (2009): 8-12.

[Handelman *et al*., 2018] Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Lee, M. J., & Asadi, H. (2018). eD octor: machine learning and the future of medicine. Journal of internal medicine, 284(6), 603-619.

[Harley et Sparkman, 2019] Harley, J. B., & Sparkman, D. (2019, May). Machine learning and NDE: Past, present, and future. In AIP conference proceedings (Vol. 2102, No. 1, p. 090001). AIP Publishing LLC.

[Hasan et Boris, 2006] Hasan, M., & Boris, F. (2006). Svm: Machines à vecteurs de support ou séparateurs à vastes marges. Rapport technique, Versailles St Quentin, France. Cité, 64.

[Hasnain et al., 2020] Hasnain, M., Pasha, M. F., Ghani, I., Imran, M., Alzahrani, M. Y., & Budiarto, R. (2020). Evaluating trust prediction and confusion matrix measures for web services ranking. IEEE Access, 8, 90847-90861.

[Hassel et Ridout, 2018] Hassel, S., & Ridout, N. (2018). An investigation of first-year students' and lecturers' expectations of university education. Frontiers in psychology, 8, 2218.

[Henrique et al., 2018] Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2018). Stock price prediction using support vector regression on daily and up to the minute prices. The Journal of finance and data science, 4(3), 183-201.

[Herodotou et al., 2019] Herodotou, C., Hlosta, M., Boroowa, A., Rienties, B., Zdrahal, Z., & Mangafa, C. (2019). Empowering online teachers through predictive learning analytics. British Journal of Educational Technology, 50(6), 3064-3079.

[Hew et al., 2020] Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. Computers & Education, 145, 103724.

[Hidayat et al., 2020] Hidayat, N., Wardoyo, R., Azhari, S., & Surjono, H. D. (2020). Enhanced Performance of the Automatic Learning Style Detection Model using a Combination of Modified K-Means Algorithm and Naive Bayesian. International Journal of Advanced Computer Science and Applications, 11, 638-48.

[Hmedna et al., 2020] Hmedna, B., El Mezouary, A., & Baz, O. (2020). A predictive model for the identification of learning styles in MOOC environments. Cluster Computing, 23(2), 1303-1328.

[Hofmann, 2001] Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42, 177–196.

[Hossain et Mahmood, 2020] Hossain, M. S., & Mahmood, H. (2020). Short-term photovoltaic power forecasting using an LSTM neural network and synthetic weather forecast. IEEE Access, 8, 172524-172533.

[Hubalovsky et al., 2019] Hubalovsky, S., Hubalovska, M., & Musilek, M. (2019). Assessment of the influence of adaptive E-learning on learning effectiveness of primary school pupils. Computers in Human Behavior, 92, 691-705.

[Hussain et al., 2018] Hussain, S., Atallah, R., Kamsin, A., & Hazarika, J. (2018, April). Classification, clustering and association rule mining in educational datasets using data mining tools: A case study. In Computer Science On-line Conference (pp. 196-211). Springer, Cham.

[Hussain et al., 2019] Hussain, M., Zhu, W., Zhang, W., Abidi, S. M. R., & Ali, S. (2019). Using machine learning to predict student difficulties from learning session data. Artificial Intelligence Review, 52(1), 381-407.

[Hussain et Prieto, 2016] Hussain, K., & Prieto, E. (2016). Big data in the finance and insurance sectors. In New horizons for a data-driven economy (pp. 209-223). Springer, Cham.

[Imani et Montazer, 2019] Imani, M., & Montazer, G. A. (2019). A survey of emotion recognition methods with emphasis on E-Learning environments. Journal of Network and Computer Applications, 147, 102423.

[Ismael et Hefny, 2020] Ismael, S. A. A., Mohammed, A., & Hefny, H. (2020). An enhanced deep learning approach for brain cancer MRI images classification using residual networks. Artificial intelligence in medicine, 102, 101779.

[Jiang et al., 2020] Jiang, Z. L., Guo, N., Jin, Y., Lv, J., Wu, Y., Liu, Z., Frang, J., Yiu, S.M., & Wang, X. (2020). Efficient two-party privacy-preserving collaborative k-means clustering protocol supporting both storage and computation outsourcing. Information Sciences, 518, 168-180.

[Jivet et al., 2018] Jivet, I., Scheffel, M., Specht, M., & Drachsler, H. (2018, March). License to evaluate: Preparing learning analytics dashboards for educational practice. In Proceedings of the 8th international conference on learning analytics and knowledge (pp. 31-40).

[Joshi et al., 2015] Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. British journal of applied science & technology, 7(4), 396.

[Juba et Le, 2019] Juba, B., & Le, H. S. (2019, July). Precision-recall versus accuracy and the role of large data sets. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 4039-4048).

[Julia et al., 2021] Julia, K., & Marco, K. (2021). Educational scalability in MOOCs: Analysing instructional designs to find best practices. Computers & Education, 161, 104054.

[Kaabi et al., 2020] Kaabi, K., Essalmi, F., Jemni, M., & Qaffas, A. A. (2020, February). Personalization of MOOCs for increasing the retention rate of learners. In 2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies" (OCTA) (pp. 1-5). IEEE.

[Kadhim, 2019] Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. Artificial Intelligence Review, 52(1), 273-292.

[Kaelbling et Moore, 1996] Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. Journal of artificial intelligence research, 4, 237-285.

[Kangalgil et Özgül, 2018] Kangalgil, M., & Özgül, F. (2018). Use of Feedback in Physical Education and Sports Lessons for Student Point of View. Universal Journal of Educational Research, 6(6), 1235-1242.

[Kaplan et Haenlein, 2016] Kaplan, A. M., & Haenlein, M. (2016). Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster. Business horizons, 59(4), 441-450.

[Karga et Satratzemi, 2018] Karga, S., & Satratzemi, M. (2018). A hybrid recommender system integrated into LAMS for learning designers. Education and Information Technologies, 23(3), 1297-1329.

[Kassambara, 2017] Kassambara, A. (2017). Practical guide to cluster analysis in R: Unsupervised machine learning (Vol. 1). STHDA.

[Kentnor, 2015] Kentnor, H. E. (2015). Distance education and the evolution of online learning in the United States. Curriculum and teaching dialogue, 17(1), 21-34.

[Kharroubi, 2002] Kharroubi, J. (2002). Etude de techniques de classement" Machines à vecteurs supports" pour la vérification automatique du locuteur (Doctoral dissertation, Télécom ParisTech, France).

[Kim et al., 2016] Kim, D., Park, Y., Yoon, M., & Jo, I. H. (2016). Toward evidence-based learning analytics: Using proxy variables to improve asynchronous online discussion environments. The Internet and Higher Education, 30, 30-43.

[Kim et al., 2019] Kim, D., Seo, D., Cho, S., & Kang, P. (2019). Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. Information Sciences, 477, 15-29.

[Kizilcec et al., 2017] Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. Computers & education, 104, 18-33.

[Klašnja et al., 2018] Klašnja-Milićević, A., Vesin, B., & Ivanović, M. (2018). Social tagging strategy for enhancing e-learning experience. Computers & Education, 118, 166-181.

[Klašnja-Milićević et al., 2017] Klašnja-Milićević, A., Vesin, B., Ivanović, M., Budimac, Z., & Jain, L. C. (2017). Recommender systems in e-learning environments. In E-Learning systems (pp. 51-75). Springer, Cham.

[Kohli et al., 2021] Kohli, S., Godwin, G. T., & Urolagin, S. (2021). Sales prediction using linear and KNN regression. In Advances in Machine Learning and Computational Intelligence (pp. 321-329). Springer, Singapore.

[Kohonen, 1982] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological cybernetics, 43(1), 59-69.

[Kokoç et Altun, 2021] Kokoç, M., & Altun, A. (2021). Effects of learner interaction with learning dashboards on academic performance in an e-learning environment. Behaviour & Information Technology, 40(2), 161-175.

[Kori et al., 2016] Kori, K., Pedaste, M., Altin, H., Tõnisson, E., & Palts, T. (2016). Factors that influence students' motivation to start and to continue studying information technology in Estonia. IEEE Transactions on Education, 59(4), 255-262.

[Krithika et Lakshmi Priya, 2016] Krithika, L. B., & Lakshmi Priya GG (2016). Student emotion recognition system (SERS) for e-learning improvement based on learner concentration metric. Procedia Computer Science, 85, 767-776.

[Kuncheva, 2014] Kuncheva, L. I. (2014). Combining pattern classifiers: methods and algorithms. John Wiley & Sons.

[Lara et Pamplona, 2020] Lara, J. A., Aljawarneh, S., & Pamplona, S. (2020). Special issue on the current trends in E-learning Assessment. Journal of Computing in Higher Education, 32(1), 1-8.

[Laureano et al., 2020] Laureano, L. B., Sison, A. M., & Medina, R. P. (2020). Affinity propagation SMOTE approach for imbalanced dataset used in predicting student at risk of low performance. International Journal of Advanced Trends in Computer Science and Engineering, 9(4), 5066-5070.

[Laqrichi, 2015] Laqrichi, S. (2015). Approche pour la construction de modèles d'estimation réaliste de l'effort/coût de projet dans un environnement incertain: application au domaine du développement logiciel (Doctoral dissertation, Ecole nationale des Mines d'Albi-Carmaux, France).

[Lebis, 2019] Lebis, A. (2019). Capitaliser les processus d'analyse de traces d'apprentissage: modélisation ontologique & assistance à la réutilisation (Doctoral dissertation, Sorbonne université).

[Lemay et al., 2021] Lemay, D. J., Baek, C., & Doleck, T. (2021). Comparison of learning analytics and educational data mining: A topic modeling approach. Computers and Education: Artificial Intelligence, 2, 100016.

[Lever et Altman, 2016] Lever, J., Krzywinski, M., & Altman, N. (2016). Logistic regression: Regression can be used on categorical responses to estimate probabilities and to classify. Nature Methods, 13(7), 541-543.

[Li et al., 2017] Li, H., Ogata, H., Tsuchiya, T., Suzuki, Y., Uchida, S., Ohashi, H., & Konomi, S. I. (2017, January). Using learning analytics to support computer-assisted language learning. In 25th International Conference on Computers in Education, ICCE 2017 (pp. 908-913). Asia-Pacific Society for Computers in Education.

[Linjawi et Alfadda, 2018] Linjawi, A. I., & Alfadda, L. S. (2018). Students' perception, attitudes, and readiness toward online learning in dental education in Saudi Arabia: a cohort study. Advances in medical education and practice, 9, 855.

[Lipton et al., 2014] Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). Thresholding classifiers to maximize F1 score. arXiv preprint arXiv:1402.1892.

[Lithio et Maitra, 2018] Lithio, A., & Maitra, R. (2018). An efficient k-means-type algorithm for clustering datasets with incomplete records. Statistical Analysis and Data Mining: The ASA Data Science Journal, 11(6), 296-311.

[Liu et al., 2008] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In 2008 eighth IEEE international conference on data mining (pp. 413-422). IEEE.

[Lopez-Bernal et al., 2021] Lopez-Bernal, D., Balderas, D., Ponce, P., & Molina, A. (2021). Education 4.0: teaching the basics of KNN, LDA and simple perceptron algorithms for binary classification problems. Future Internet, 13(8), 193.

[Lung-Guang, 2019] Lung-Guang, N. (2019). Decision-making determinants of students participating in MOOCs: Merging the theory of planned behavior and self-regulated learning model. Computers & Education, 134, 50-62.

[Luque et al., 2019] Luque, A., Carrasco, A., Martín, A., & de Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 91, 216-231.

[Lévy, 2021] Lévy, M. (2021). Chapitre 1. La data est-elle vraiment le pétrole du XXIe siècle ?. Dans : M. Lévy, Sortez vos données du frigo: Une entreprise performante avec la Data et l'iA (pp. 23-26). Paris: Dunod.

[Madhulatha, 2012] Madhulatha, T. S. (2012). An overview on clustering methods. arXiv preprint arXiv:1205.1117.

[Mahesh, 2020] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR), 9, 381-386.

[Makkar et al., 2020] Makkar, A., Garg, S., Kumar, N., Hossain, M. S., Ghoneim, A., & Alrashoud, M. (2020). An efficient spam detection technique for IoT devices using machine learning. IEEE Transactions on Industrial Informatics, 17(2), 903-912.

[Malhotra et Rishi, 2021] Malhotra, D., & Rishi, O. (2021). An intelligent approach to design of E-Commerce metasearch and ranking system using next-generation big data analytics. Journal of King Saud University-Computer and Information Sciences, 33(2), 183-194.

[Margaryan et al., 2015] Margaryan, A., Bianco, M., & Littlejohn, A. (2015). Instructional quality of massive open online courses (MOOCs). Computers & Education, 80, 77-83.

[Maulud et Abdulazeez, 2020] Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(4), 140-147.

[McGillicuddy, 2020] McGillicuddy, K. T. (2020). Gaming for a Grade: How Goal Achievement and Causality Orientations Impact the Efficacy of Games for Classroom Engagement, Academic Achievement, and Learner Satisfaction (Doctoral dissertation, Université du Connecticut, Storrs, Etats Unis).

[Mekonnen et al., 2019] Mekonnen, Y., Namuduri, S., Burton, L., Sarwat, A., & Bhansali, S. (2019). Machine learning techniques in wireless sensor network based precision agriculture. Journal of the Electrochemical Society, 167(3), 037522.

[Melesko et Kurilovas, 2018] Melesko, J., & Kurilovas, E. (2018, June). Semantic technologies in e-learning: Learning analytics and artificial neural networks in personalised learning systems. In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics (pp. 1-7).

[Mengoni et al., 2018] Mengoni, P., Milani, A., & Li, Y. (2018, May). Clustering students interactions in eLearning systems for group elicitation. In International Conference on Computational Science and Its Applications (pp. 398-413). Springer, Cham.

[Miss MOOC, 2016] Miss MOOC, (2016). Quand SNCF fait son COOC: Le MOOC incivilité. Available from https://miss-mooc.paris/2016/11/15/quand-sncf-fait-son-cooc-le-mooc-incivilites/ [Online]. [Accessed: 10-Jul-2022]

[Mitchell, 1997] Mitchell, T. M. (1997). Machine Learning. McGraw-Hill. New York, 154-200.

[Mitra, 2020] Mitra, A. (2020). Sentiment analysis using machine learning approaches (Lexicon based on movie review dataset). Journal of Ubiquitous Computing and Communication Technologies (UCCT), 2(03), 145-152.

[Montgomery et al., 2019] Montgomery, A. P., Mousavi, A., Carbonaro, M., Hayward, D. V., & Dunn, W. (2019). Using learning analytics to explore self-regulated learning in flipped blended learning music teacher education. British Journal of Educational Technology, 50(1), 114-127.

[Moubayed et al., 2020] Moubayed, A., Injadat, M., Shami, A., & Lutfiyya, H. (2020). Student engagement level in an e-learning environment: Clustering using k-means. American Journal of Distance Education, 34(2), 137-156.

[Mourali et al., 2020a] Mourali, Y., Agrebi, M., Ezzedine, H., Farhat, R., Jemni, M., & Abed, M. (2020). A Review On E-learning: Perspectives And Challenges. ICIW 2020: The Fifteenth International Conference on Internet and Web Applications and Services, Portugal, 1-7.

[Mourali et al., 2020b] Mourali, Y., Farhat, R., Jemni, M., & Ezzedine, H. (2020). E-learning and machine learning–A Look at the future of education technologies. Multidimensionality of research for sustainable development, Cambridge Scholars Publishing, 2020.

[Mourali et al., 2021a] Mourali, Y., Agrebi, M., Farhat, R., Ezzedine, H., & Jemni, M. (2021). Learning Analytics Metrics into Online Course's Critical Success Factors. In WorldCIST (2) (pp. 161-170).

[Mourali et al., 2021b] Mourali, Y., Farhat, R., Agrebi, M., Jemni, M., Kolski, C., & Ezzedine, H. (2021, December). An educational decision support system: case of learners clustering. In 2021 8th International Conference on ICT & Accessibility (ICTA) (pp. 1-3). IEEE.

[Mubarak et Ahmed, 2021] Mubarak, A. A., Cao, H., & Ahmed, S. A. (2021). Predictive learning analytics using deep learning model in MOOCs' courses videos. Education and Information Technologies, 26(1), 371-392.

[Mydhili et al., 2020] Mydhili, S. K., Periyanayagi, S., Baskar, S., Shakeel, P. M., & Hariharan, P. R. (2020). Machine learning based multi scale parallel K-means++ clustering for cloud assisted internet of things. Peer-to-Peer Networking and Applications, 13(6), 2023-2035.

[Mystakidis et al., 2021] Mystakidis, S., Berki, E., & Valtanen, J. P. (2021). Deep and meaningful e-learning with social virtual reality environments in higher education: A systematic literature review. Applied Sciences, 11(5), 2412.

[Nasteski, 2017] Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons, 4, 51-62.

[Nawrin et al., 2017] Nawrin, S., Rahman, M. R., & Akhter, S. (2017). Exploreing k-means with internal validity indexes for data clustering in traffic management system. International Journal of Advanced Computer Science and Applications, 8(3), 264-272.

[Oliver et al., 2018] Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., & Goodfellow, I. (2018). Realistic evaluation of deep semi-supervised learning algorithms. Advances in neural information processing systems. Proceedings of the 32nd International Conference on Neural Information Processing Systems, 3239–3250. Montréal, Canada. Red Hook, NY, USA: Curran Associates Inc.

[Ostrovsky et al., 2013] Ostrovsky, R., Rabani, Y., Schulman, L. J., & Swamy, C. (2013). The effectiveness of Lloyd-type methods for the k-means problem. Journal of the ACM (JACM), 59(6), 1-22.

[Ouadoud et al., 2021] Ouadoud, M., Rida, N., & Chafiq, T. (2021). Overview of E-learning Platforms for Teaching and Learning. Int. J. Recent Contributions Eng. Sci. IT, 9(1), 50-70.

[Padilla et al., 2020] Padilla, R., Netto, S. L., & Da Silva, E. A. (2020, July). A survey on performance metrics for object-detection algorithms. In 2020 international conference on systems, signals and image processing (IWSSIP) (pp. 237-242). IEEE.

[Page et al., 2014] Page, J. T., Liechty, Z. S., Huynh, M. D., & Udall, J. A. (2014). BamBam: genome sequence analysis tools for biologists. BMC Research Notes, 7(1), 1-5.

[Pardos et Kao, 2015] Pardos, Z. A., & Kao, K. (2015, March). moocRP: An open-source analytics platform. In Proceedings of the Second (2015) ACM conference on learning@ scale (pp. 103-110).

[Paullada et al., 2021] Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns, 2(11), 100336.

[Peng et al., 2002] Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), 3-14.

[Phaladisailoed et Numnonda, 2018] Phaladisailoed, T., & Numnonda, T. (2018, July). Machine learning models comparison for bitcoin price prediction. In 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE) (pp. 506-511). IEEE.

[Pham et al., 2019] Pham, L., Limbu, Y. B., Bui, T. K., Nguyen, H. T., & Pham, H. T. (2019). Does e-learning service quality influence e-learning student satisfaction and loyalty? Evidence from Vietnam. International Journal of Educational Technology in Higher Education, 16(1), 1-26.

[Popchev et Orozova, 2019] Popchev, I. P., & Orozova, D. A. (2019). Towards big data analytics in the e-learning space. Cybernetics and information technologies, 19(3), 16-24.

[Portugal et al., 2018] Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. Expert Systems with Applications, 97, 205-227.

[Qiu et al., 2016] Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing, 2016(1), 1-16.

[Raj, 2021] Raj, S. (2021). Prioritization of e-learners activities using principal component analysis method. International Journal of Information Technology, 13(6), 2439-2451.

[Rajkomar et Kohane, 2019] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358.

[Ramesh et al., 2013] Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., & Getoor, L. (2013, December). Modeling learner engagement in MOOCs using probabilistic soft logic. In NIPS workshop on data driven education (Vol. 21, p. 62).

[Raschka et Nolet, 2020] Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. Information, 11(4), 193.

[Rathi et al., 2018] Rathi, M., Malik, A., Varshney, D., Sharma, R., & Mendiratta, S. (2018, August). Sentiment analysis of tweets using machine learning approach. In 2018 Eleventh international conference on contemporary computing (IC3) (pp. 1-3). IEEE.

[Ray, 2019] Ray, S. (2019, February). A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 35-39). IEEE.

[Refaeilzadeh and Liu, 2009] Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. Encyclopedia of database systems, 5, 532-538.

[Reid, 2012] Reid, L. F. (2012). Redesigning a large lecture course for student engagement: Process and outcomes. The Canadian Journal for the Scholarship of Teaching and Learning, 3(2). https://doi.org/10.5206/cjsotl-rcacea.2012.2.5

[Reinert, 1976] Reinert, H. (1976). One picture is worth a thousand words? Not necessarily!. Modern Language Journal, 160-168.

[Reinforcement Learning, 2017] Reinforcement Learning (2017). In Wikipedia. https://en.wikipedia.org/wiki/Reinforcement_learning

[Renault Group, 2018] Renault Group, (2018). Les véhicules électriques et la mobilité : un MOOC pour se forger une opinion. Available from https://www.renaultgroup.com/news-onair/actualites/mobilites-et-les-vehicules-electriques-un-mooc-pour-se-forger-une-opinion/ [Online]. [Accessed: 10-Jul-2022]

[Risi et Preuss, 2020] Risi, S., & Preuss, M. (2020). From chess and atari to starcraft and beyond: How game ai is driving the world of ai. KI-Künstliche Intelligenz, 34(1), 7-17.

[Rodrigues et al., 2018] Rodrigues, M. W., Isotani, S., & Zarate, L. E. (2018). Educational Data Mining: A review of evaluation process in the e-learning. Telematics and Informatics, 35(6), 1701-1717.

[Romero et Ventura, 2020] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355.

[Safaei-Farouji et al., 2022] Safaei-Farouji, M., Band, S. S., & Mosavi, A. (2022). Oil Family Typing Using a Hybrid Model of Self-Organizing Maps and Artificial Neural Networks. ACS omega, 7(14), 11578–11586.

[Sagheer et al., 2019] Sagheer, A., Zidan, M., & Abdelsamea, M. M. (2019). A novel autonomous perceptron model for pattern classification applications. Entropy, 21(8), 763.

[Salloum et al., 2020] Salloum, S. A., Alshurideh, M., Elnagar, A., & Shaalan, K. (2020, April). Mining in educational data: review and future directions. In Joint European-US Workshop on Applications of Invariance in Computer Vision (pp. 92-102). Springer, Cham.

[Salminen et al., 2019] Salminen, J., Yoganathan, V., Corporan, J., Jansen, B. J., & Jung, S. G. (2019). Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type. Journal of Business Research, 101, 203-217.

[Samuelsen et al., 2019] Samuelsen, J., Chen, W., & Wasson, B. (2019). Integrating multiple data sources for learning analytics—review of literature. Research and Practice in Technology Enhanced Learning, 14(1), 1-20.

[Sangrà et al., 2012] Sangrà, A., Vlachopoulos, D., & Cabrera, N. (2012). Building an inclusive definition of e-learning: An approach to the conceptual framework. International Review of Research in Open and Distributed Learning, 13(2), 145-159.

[Sarker, 2021] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2(3), 1-21.

[Sasaki, 2007] Sasaki, Y. (2007). The truth of the F-measure. Teach tutor mater, 1(5), 1-5.

[Sathish et al., 2020] Sathish, T., Rangarajan, S., Muthuram, A., & Kumar, R. P. (2020). Analysis and modelling of dissimilar materials welding based on K-nearest neighbour predictor. Materials Today: Proceedings, 21, 108-112.

[Saxena et al., 2017] Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., ... & Lin, C. T. (2017). A review of clustering techniques and developments. Neurocomputing, 267, 664-681.

[Schrider et Kern, 2018] Schrider, D. R., & Kern, A. D. (2018). Supervised machine learning for population genetics: a new paradigm. Trends in Genetics, 34(4), 301-312.

[Scikit-learn] Scikit-learn. Clustering. Available from https://scikit-learn.org/stable/modules/clustering.html [Online]. [Accessed: 25-Aug-2022]

[Senders et al., 2018] Senders, J. T., Staples, P. C., Karhade, A. V., Zaki, M. M., Gormley, W. B., Broekman, M. L., ... & Arnaout, O. (2018). Machine learning and neurosurgical outcome prediction: a systematic review. World neurosurgery, 109, 476-486.

[Sezer et Altan, 2021] Sezer, A., & Altan, A. (2021, June). Optimization of deep learning model parameters in classification of solder paste defects. In 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) (pp. 1-6). IEEE.

[Shang et You, 2019] Shang, C., & You, F. (2019). Data analytics and machine learning for smart process manufacturing: recent advances and perspectives in the big data era. Engineering, 5(6), 1010-1016.

[Sharma et Kumar, 2017] Sharma, D., & Kumar, N. (2017). A review on machine learning algorithms, tasks and applications. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 6(10), 2278-1323.

[Sherstinsky, 2020] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.

[Simeone, 2018] Simeone, O. (2018). A brief introduction to machine learning for engineers. Foundations and Trends® in Signal Processing, 12(3-4), 200-431.

[Sinaga et Yang, 2020] Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE access, 8, 80716-80727.

[Singh et Thurman, 2019] Singh, V., & Thurman, A. (2019). How many ways can we define online learning? A systematic literature review of definitions of online learning (1988-2018). American Journal of Distance Education, 33(4), 289-306.

[Soykan et Kanbul, 2018] Soykan, F., & Kanbul, S. (2018). Analysing K12 students' self-efficacy regarding coding education. TEM Journal, 7(1), 182.

[Srinivasan et Arunasalam, 2013] Srinivasan, U., & Arunasalam, B. (2013). Leveraging big data analytics to reduce healthcare costs. IT professional, 15(6), 21-28.

[Stilgoe, 2018] Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. Social studies of science, 48(1), 25-56. https://doi.org/10.1177/0306312717741687

[Strogatz, 2018] Strogatz, S. (2018). One giant step for a chess-playing machine. New York Times, 1-6.

[Sudhahar et al., 2015] Sudhahar, S., Veltri, G. A., & Cristianini, N. (2015). Automated analysis of the US presidential elections using Big Data and network analysis. Big Data & Society, 2(1), 2053951715572916.

[Sui et al., 2021] Sui, X., He, S., Vilsen, S. B., Meng, J., Teodorescu, R., & Stroe, D. I. (2021). A review of non-probabilistic machine learning-based state of health estimation techniques for Lithium-ion battery. Applied Energy, 300, 117346.

[Takahashi et al., 2022] Takahashi, K., Yamamoto, K., Kuchiba, A., & Koyama, T. (2022). Confidence interval for micro-averaged F1 and macro-averaged F1 scores. Applied Intelligence, 52(5), 4961-4972.

[Tang et al., 2017] Tang, J., Wang, D., Zhang, Z., He, L., Xin, J., & Xu, Y. (2017). Weed identification based on K-means feature learning combined with convolutional neural network. Computers and Electronics in Agriculture, 135, 63-70.

[Tao et al., 2016] Tao, L. J., Hong, L. Y., & Yan, H. (2016, July). The improvement and application of a K-means clustering algorithm. In 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) (pp. 93-96). IEEE.

[Taunk et al., 2019] Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019, May). A brief review of nearest neighbor algorithm for learning and classification. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 1255-1260). IEEE.

[Timbal, 2019] Timbal, M. A. (2019). Analysis of Student-at-Risk of Dropping out (SARDO) Using Decision Tree: An Intelligent Predictive Model for Reduction. International Journal of Machine Learning and Computing, 9(3), 273-278.

[Torrent-Fontbona et al., 2012] Torrent-Fontbona, F., Muñoz Solà, V., & López Ibáñez, B. (2012). Decision Support Methods for Global Optomization (Doctoral dissertation, University of Girona, Girona, Spain).

[Toti et al., 2020] Toti, D., Capuano, N., Campos, F., Dantas, M., Neves, F., & Caballé, S. (2020, October). Detection of student engagement in e-learning systems based on semantic analysis and machine learning. In International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (pp. 211-223). Springer, Cham.

[Trehan et Joshi, 2018] Trehan, S., & Joshi, R. (2018). Building and evaluating logistic regression models for explaining the choice to adopt MOOCs in India. International Journal of Education and Development using ICT, 14(1), 33-51.

[Tsolou et al., 2021] Tsolou, O., Babalis, T., & Tsoli, K. (2021). The impact of COVID-19 pandemic on education: social exclusion and dropping out of school. Creative Education, 12(03), 529.

[Uthayakumar et al., 2020] Uthayakumar, J., Vengattaraman, T., & Dhavachelvan, P. (2020). Swarm intelligence based classification rule induction (CRI) framework for qualitative and quantitative approach: An application of bankruptcy prediction and credit risk analysis. Journal of King Saud University-Computer and Information Sciences, 32(6), 647-657.

[Viberg et al., 2020] Viberg, O., Khalil, M., & Baars, M. (2020, March). Self-regulated learning and learning analytics in online learning environments: A review of empirical research.

In Proceedings of the tenth international conference on learning analytics & knowledge (pp. 524-533).

[Von Luxburg, 2007] Von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and computing, 17(4), 395-416.

[Wang et al., 2019] Wang, J., Gao, Y., Wang, K., Sangaiah, A. K., & Lim, S. J. (2019). An affinity propagation-based self-adaptive clustering method for wireless sensor networks. Sensors, 19(11), 2579.

[Wong et al., 2018] Wong, E. Y., Kwong, T., & Pegrum, M. (2018). Learning on mobile augmented reality trails of integrity and ethics. Research and Practice in Technology Enhanced Learning, 13(1), 1-20.

[Wong et al., 2019] Wong, J., Baars, M., Davis, D., Van Der Zee, T., Houben, G. J., & Paas, F. (2019). Supporting self-regulated learning in online learning environments and MOOCs: A systematic review. International Journal of Human–Computer Interaction, 35(4-5), 356-373.

[Wu, 2021] Wu, B. (2021, August). K-means clustering algorithm and Python implementation. In 2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE) (pp. 55-59). IEEE.

[Yang et al., 2018] Yang, S. J., Lu, O. H., Huang, A. Y., Huang, J. C., Ogata, H., & Lin, A. J. (2018). Predicting students' academic performance using multiple linear regression and principal component analysis. Journal of Information Processing, 26, 170-176.

[Yue et al., 2010] Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC medical informatics and decision making, 10(1), 1-7.

[Yue et al., 2018] Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. Designs, 2(2), 13.

[Yue et al., 2019] Yu, C. H., Wu, J., & Liu, A. C. (2019). Predicting learning outcomes with MOOC clickstreams. Education sciences, 9(2), 104.

[Zhang et al., 2019] Zhang, H., Li, Z., Shahriar, H., Tao, L., Bhattacharya, P., & Qian, Y. (2019, July). Improving prediction accuracy for logistic regression on imbalanced datasets. In 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC) (Vol. 1, pp. 918-919). IEEE.

[Zhu et al., 2019] Zhu, G., Wu, Z., Wang, Y., Cao, S., & Cao, J. (2019). Online purchase decisions for tourism e-commerce. Electronic Commerce Research and Applications, 38, 100887.

[Ünlü et Xanthopoulos, 2019] Ünlü, R., & Xanthopoulos, P. (2019). Estimating the number of clusters in a dataset via consensus clustering. Expert Systems with Applications, 125, 33-39.

Annexe 1 : La méthodologie CRISP (Cross-Industry Standard Process)

Introduction

Dans le cadre de cette thèse nous avons adopté la méthodologie des sciences des données CRISP pour développer le système d'aide à l'évaluation et l'amélioration des contenus éducatifs en ligne combinant les deux approches MALEA et ACSP.

Les critères de choix de la méthodologie CRISP

Initialement connue sous le nom de CRISP-DM, la méthode CRISP, a été conçue en 1998 avec le soutien d'IBM. Elle a été développée et publiée officiellement en 2000. En 2003, elle est devenue un standard de facto. Elle a été conçue, au départ spécifiquement pour l'exploration de données. Cependant, elle est suffisamment flexible pour pouvoir être appliquée à tout projet analytique, qu'il s'agisse de la science des données, l'apprentissage automatique, l'analyse prédictive, etc. Cette méthodologie a la particularité d'adopter une démarche cyclique et itérative permettant une meilleure appréhension des spécificités de chaque projet. Ainsi, la méthodologie CRISP est un cadre de référence pour la réalisation des projets de science des données. Elle propose une approche structurée avec des tâches bien définies pour aider à planifier, organiser et mettre en œuvre son projet. Cette méthode est agile et itérative dans laquelle chaque itération apporte une connaissance supplémentaire qui permet de mieux aborder l'itération suivante.

Les étapes du processus CRISP

La méthode CRISP se décompose en 6 étapes allant de la compréhension du problème métier au déploiement et la mise en production. La Figure 6.1 fournit une représentation visuelle du processus et montre les boucles de rétroaction, qui facilitent sa flexibilité.

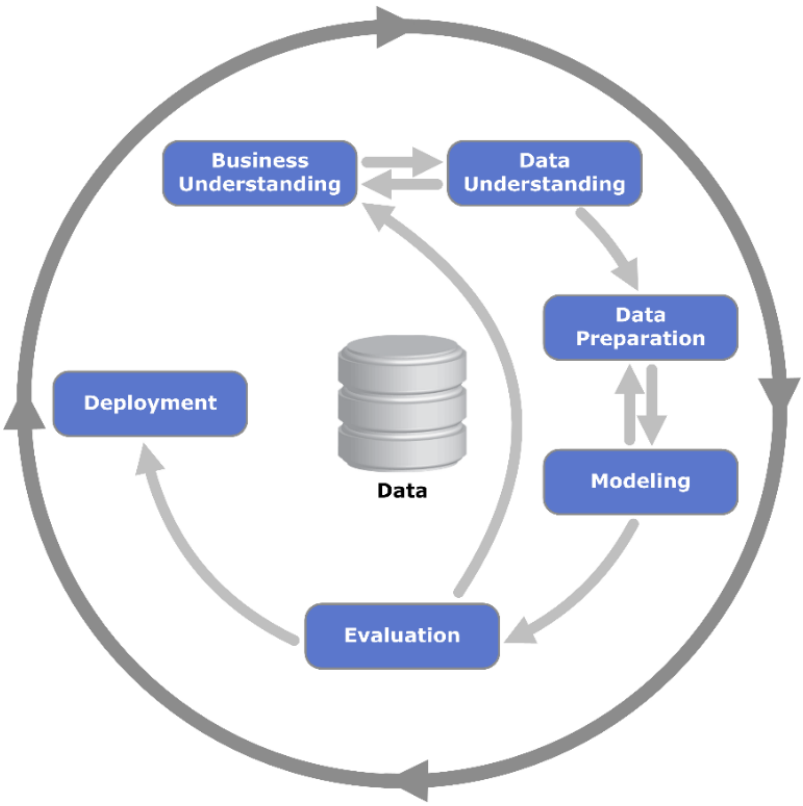

Figure 6.1 Illustration de la méthode CRISP

Etape 1 : Compréhension de la problématique

La première étape du processus CRISP est fondamentale. Elle se concentre sur la compréhension des objectifs et des exigences du projet. Son but consiste à identifier les objectifs métiers afin de pouvoir les traduire en objectifs d'analyse. Il est important de déterminer les bénéfices que nous souhaitons tirer de l'analyse des données et clarifier les problèmes qui peuvent être résolus (via les méthodes de l'analyse). A la fin de cette étape, un plan de projet devrait être élaboré dans lequel la technologie et les outils à utiliser sont définis et les différentes phases du projet sont détaillées.

Etape 2 : Compréhension des données

Cette étape commence par le recensement des données existantes. Il est crucial de déterminer si les données disponibles sont adéquates pour répondre aux besoins d'analyse. Examiner les données, décrire leurs caractéristiques, les explorer, visualiser et interpréter sont des tâches qui aident à la détection des informations pertinentes pouvant servir à la construction des modèles. La qualité des données est vérifiée par l'identification des erreurs, des valeurs manquantes, etc.

Etape 3 : Préparation des données

Cette étape est la plus longue d'un projet de science de données. Elle monopolise généralement 80% du temps consacré à l'ensemble du projet. La préparation de données couvre toutes les activités pour construire l'ensemble de données finales qui seront transmises aux outils de l'analyse. Dans cette étape, il sera question de sélection, d'intégration, de transformation et de formatage des données sous des formes appropriées. Si la labélisation est nécessaire, elle est exécutée. C'est à ce stade que s'opère la répartition des données sous différents ensembles comme le test et l'entrainement. Cette étape comprend, aussi, le nettoyage des données pour éliminer le bruit et l'inconsistance, le traitement des valeurs manquantes et des points atypiques, etc.

Etape 4 : Modélisation

C'est la phase de construction et d'évaluation de divers modèles basés sur une variété de techniques de modélisation disponibles. La modélisation inclut le choix des techniques et des algorithmes de modélisation, la génération des tests, la création et l'évaluation des modèles**.** Parfois, il y aura une phase de va et vient entre la préparation de données et la modélisation pour pouvoir utiliser certains algorithmes. L'étape de la modélisation va générer, souvent, plusieurs modèles qui répondent à la problématique. Une évaluation préalable de ces modèles est souhaitable.

Etape 5 : Evaluation

Tandis que la tâche d'évaluation du modèle de la phase de modélisation se concentre sur l'évaluation technique du modèle en lui-même, la phase d'évaluation examine plus largement quel modèle ou ensemble de modèles présentent les meilleures performances.

Etape 6 : Déploiement

Cette étape consiste à intégrer le modèle obtenu au processus de prise de décision. Ainsi, le déploiement peut aller, selon les objectifs, de la génération d'un rapport décrivant les connaissances obtenues jusqu'à la mise en place d'une application. Cette étape ne signifie pas nécessairement la fin du processus car rajouter de nouvelles données pourrait améliorer le modèle.

Annexe 2 : Liste des publications

Articles et communications publiés :

Chapitre d'ouvrage :

1) Mourali, Y., Farhat, R., Jemni, M., & Ezzedine, H. (2020). E-learning and machine learning–A Look at the future of education technologies. Multidimensionality of research for sustainable development, Cambridge Scholars Publishing, 2020.
URL: https://hal.archives-ouvertes.fr/hal-03382074/

Communications dans des congrès :

1) Farhat, R., Mourali, Y., Jemni, M., & Ezzedine, H. (2020, February). An overview of Machine Learning Technologies and their use in E-learning. In 2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA) (pp. 1-4). IEEE.
URL: http://dx.doi.org/10.1109/octa49274.2020.9151758

2) Mourali, Y., Agrebi, M., Ezzedine, H., Farhat, R., Jemni, M., & Abed, M. (2020). A Review On E-learning: Perspectives And Challenges. ICIW 2020 The Fifteenth International Conference on Internet and Web Applications and Services, Lisbone, Portugal.
URL:
http://www.thinkmind.org/index.php?view=article&articleid=iciw_2020_1_10_20010

3) Mourali, Y., Agrebi, M., Farhat, R., Ezzedine, H., & Jemni, M. (2021). Learning Analytics Metrics into Online Course's Critical Success Factors. In WorldCIST (2) (pp. 161-170).
URL:https://books.google.tn/books?hl=fr&lr=&id=jD8mEAAAQBAJ&oi=fnd&pg=PA161&dq=yosra+mourali&ots=caHgEP04Yl&sig=BD2TTUNTQiIlWxG15VoK4TStEDc&redir_esc=y#v=onepage&q=yosra%20mourali&f=false

4) Mourali, Y., Farhat, R., Agrebi, M., Jemni, M., Kolski, C., & Ezzedine, H. (2021, December). An educational decision support system: case of learners clustering. In 2021 8th International Conference on ICT & Accessibility (ICTA) (pp. 1-3). IEEE.
URl: https://ieeexplore.ieee.org/abstract/document/9809425

Article en soumission dans un journal :

"Evaluation of online educational content based on learning experience analysis"